

The end of innocence?

Biologists must become more aware that their work could be abused to develop weapons of mass destruction. But it should not have taken the disturbing events of recent weeks to bring this debate to the fore.

In suggesting that biologists must regulate their work to prevent the free flow of information to would-be bioweapons developers, George Poste will win himself few friends at the lab bench. Yet biologists would do well to ponder on the comments of this adviser to the US Department of Defense, and former senior executive in the drugs industry, made at a conference in London (see page 237).

Poste accuses biologists of naivety in failing to consider the possibility that their research findings could be used for malign purposes. *Nature* expressed similar sentiments in an editorial published in May this year (411, 223; 2001).

But the difficulty is how to respond to the fact that data generated in legitimate projects could have applications in bioweapons development. The anguished reactions to some of Poste's suggestions — such as the vetting of manuscripts to determine whether 'sensitive' findings should be withheld from the open literature — show that there are no simple answers. Although such ideas might not shock cryptographers, who are used to dealing with issues of national security, this is uncharted territory for most biologists.

Scientific societies and research agencies should now rise to the challenge by holding meetings to debate the issues. Poste points to the example of the 1975 Asilomar Conference on Recombinant DNA Molecules, which considered the risks of recombinant DNA research at a time when US biologists were observing a voluntary research moratorium. Out of this came guidelines that were adopted, in modified form, by the National Institutes of Health.

The potential applications of advances in biology to the development of 'enhanced' bioweapons poses more complex problems, however, and it is unlikely that a clear consensus will emerge. It is also important to understand the nature of the threats involved. The sorts of projects that Poste is most worried about — such as the creation of 'stealth' viruses that would evade the immune system — are most probably beyond the technological capabilities of whoever is behind the current mailed anthrax attacks in the United States. It would be inappropriate, therefore, to impose stringent restrictions on such work as a reaction to the immediate terrorist threat.

Such technologies are within the reach of a well-funded state bioweapons programme, however. So, in the longer term, Poste is right to be concerned. The most useful result of the debate he wishes to stimulate would be a heightened awareness among biologists of the potential dangers, which hopefully would influence decisions on whether or not to pursue particularly risky projects in the first place.

When experimenting with biotechnologies that try to simulate and accelerate the natural evolution of traits, for instance, is it really wise to test them by heightening the efficiency of genes that confer bacterial resistance to antibiotics, as some researchers have done? The same question can be asked of work that splices genes for cytokines — proteins that manipulate immune responses — into potentially pathogenic viruses.

Some self-restraint is desirable when pursuing the frontiers of biology. Discuss. ■

Essentials for China's development

A balance of basic and applied science makes sense, provided resources are deployed wisely.

At an opening ceremony of the Chinese Science Association in September, Guangzhao Zhou, former president of the Chinese Academy of Sciences, noted some undesirable tendencies in the Chinese scientific community — a rush to get results, imitative work, a fixation on getting large numbers of publications, overstatement of scientific achievements, hesitation in engaging in academic debate and lack of interdisciplinary collaboration. Even worse, he noted, some people engage in unethical conduct such as scientific fraud, self-promotion and plagiarism.

Despite such criticisms, the development of China's research has an air of inevitability. Beijing and Shanghai are buzzing with scientific activity as the government has recruited many promising young scientists from abroad. These are mostly ethnic Chinese, whose families came from mainland China, although Tsinghua University recently appointed its first Western chair of department — in industrial engineering. Like all developing countries, China is looking to new technology for industrial applications. But it has also been far-sighted enough to pour money into fundamental research initiatives that have the makings of world-class science (see page 240).

But low salaries mean that many researchers depend greatly on grants or other rewards, such as those given for publications in major scientific journals. Reliance on frequent publication for evaluation

puts excessive pressure on researchers, even more than in the West. Scientists are too often tempted to submit reports of work that represents only a promising start. To bear fruit requires time, and it is not clear that China is willing to be patient.

The Chinese government needs to give more researchers, both academic and applied, better salaries and a firm footing so that they do not need to scramble for proof of their worthiness. Academic researchers should be encouraged to apply their research, and science should be capitalized on wherever possible. For what China needs is an industrial base that can support research that will complement the work done in universities and academy institutes. Only then can the government's burden of overseeing most scientific research be lightened. But application of science, like good academic results, also takes time.

To evaluate new technology properly will require a strong private industrial base that is in the business of long-term profitability, not state-owned companies out to impress government with flashy products and a quick profit. And to evaluate researchers for grants without depending solely on past publications, China needs to establish a wider network of external reviewers.

China's scientific development seems inexorable. It has assembled considerable human and financial resources. But unless it makes wise policies now, much of this money and talent will be wasted. ■

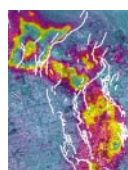




Boosted
US national vaccine
laboratory edges
closer to reality
p239



Poles apart
Habitat fears raise
questions over plans
to drill in the Arctic
p240



Core project
EarthScope gets vote
of confidence from
National Academies
p241



Final protest
Animal-rights
campaigner dies
after hunger strike
p242

Biologists urged to address risk of data aiding bioweapon design

Peter Aldhous, London

Biologists must begin a process of self-regulation for projects that have potential applications in developing bioweapons — or risk the imposition of restrictive controls from outside. That is the sobering message being delivered by a senior scientific adviser to the US defence department.

George Poste, who chairs a Department of Defense task force on bioterrorism and sits on its advisory Defense Science Board, told a pharmaceuticals-industry conference in London on 6 November that biology must “lose its innocence”. He criticized biologists for what he regards as their naivety in failing to consider malign applications of data generated in legitimate projects, and urged them to start considering how access to such data could best be regulated.

“What I think is untenable is the status quo — just allowing highly sensitive information to enter the public domain,” says Poste, a former head of research at the drugs company



Fear factor: restrictions could extend beyond work on bioweapons agents such as these anthrax spores.

SmithKline Beecham. “Equally dangerous would be a draconian legislative response.”

With the United States reeling in the face of anthrax attacks against media outlets and politicians, Poste fears that unduly restrictive legislation could be hurried through. He points to a bill that is now before Congress that

would impose sweeping restrictions on non-US citizens handling potential bioweapons agents, which has already caused alarm among biologists (see *Nature* **414**, 3–4; 2001).

In an interview with *Nature*, Poste expanded upon his concerns. He is particularly worried about projects in which viruses are engineered to evade or manipulate the immune system. For instance, earlier this year Australian researchers revealed that they had inadvertently created a super-virulent strain of mousepox in a project aimed at creating a contraceptive vaccine (see *Nature* **411**, 232–235; 2001). And gene therapists, grappling with the problems caused by immune reactions to the viral vectors they use to introduce therapeutic genes, are now designing ‘stealth’ vectors that would escape the attention of the immune system.

Such technologies could be applied to viral bioweapons with devastating effects, argues Poste. Yet he has been disappointed by the reactions of some researchers working on such projects when this possibility is raised: “They look at you first of all with a blank stare, followed by arrogant denial,” he says.

Poste believes that a larger number of projects addressing the issue of biodefence may in future have to be classified. More generally, he suggests that researchers submitting grant proposals to bodies such as the National Institutes of Health (NIH) might have to declare whether they have considered potential malign applications. For projects deemed to be particularly risky, manuscripts

CERN opens finances up for review

David Adam, London

CERN, the European Laboratory for Particle Physics near Geneva, is seeking help from outside experts as it struggles to find a way through the funding crisis triggered by cost overruns on its Large Hadron Collider (LHC) project.

At a meeting of the laboratory's finance committee on 6 November, CERN managers said they would appoint an external review board to assess future financing needs through to 2012 for both the LHC and the lab as a whole. The board will be set up later this month and should produce a preliminary report by the end of this year, with a final one to follow in June 2002.

The laboratory's effort to build the LHC, the world's most powerful proton accelerator, has gone over budget by several hundred million dollars (see *Nature* **413**, 441; 2001).

In a statement posted on its website on 16 October, CERN's director-general,

Luciano Maiani, hinted strongly that the LHC could not be completed without more money from the member states. “CERN's capacity to absorb these extra costs has been severely limited,” he wrote. “Support from the governments and funding agencies is now essential to reach a solution.”

But at the finance meeting, several delegates insisted that there would be no extra contributions from them to bail out the project.

Neil Calder, a spokesman for CERN, says that Maiani's call for support was not necessarily a request for more money. He claims that, if required, CERN could build the LHC without extra contributions but “that would not be the optimum solution”.

Meanwhile, CERN has appointed four internal ‘task forces’ to find savings from its research programmes, its staffing, its industrial links and its organizational structure.

might even be vetted, he says, with the possibility that permission to publish could be denied.

Poste also suggests that anyone wishing to access genomic sequence data for dangerous pathogens might be required to provide evidence of their accreditation with a legitimate lab — a stipulation already applied by suppliers of microbial cultures.

But Poste stresses that wider debate among biologists is needed to develop an appropriate framework for self-regulation. "These are not formal proposals," he says.

Poste's suggestions are highly controversial, given that the freedom to publish and to share data is central to the culture of modern bioscience. "We consider openness in publication as sacrosanct in our work," says Tony Fauci, director of the National Institute of Allergy and Infectious Diseases in Bethesda, Maryland. "I think that's something that we are going to have to be very careful about treading on."



George Poste:
wants to restrict
access to data.

Fauci is similarly sceptical about proposals to restrict access to genome-sequence databases. "I'm not sure how much safer the public would be" as a result of such controls, he says.

Gene therapists, who are already subject to stringent controls, view any suggestion of additional regulation with horror. "I spend 90% of my time dealing with regulatory authorities," says Malcolm Brenner of Baylor College of Medicine in Houston, Texas, president of the American Society of Gene Therapy. "Any further regulation will be the kiss of death."

Mark Wheelis, a microbiologist at the University of California, Davis, and a member of the Federation of American Scientists' working group on bioweapons, is alarmed by Poste's suggestion of classifying a greater proportion of work on defence against bioweapons. He argues that this could allow states to hide offensive bioweapons programmes under the guise of classified biodefence projects — or at least to create suspicions that they are doing so.

On the basis of his conversations with members of Congress, Fauci says he is confident that legislators will not impose draconian restrictions. But both he and Wheelis agree that biologists must become more aware of potential destructive applications of their work. They urge scientific societies to take a lead in promoting debate.

Fauci says the NIH could play a part by hosting discussion meetings, and believes that heightened awareness could lead to biologists deciding not to go ahead with certain projects. ■



Optimistic: delegates at the Marrakech meeting believe the Kyoto Protocol could soon come into force.

Accord in Morocco breathes fresh life into Kyoto Protocol

Quirin Schiermeier

The Kyoto Protocol on climate change now has an official rulebook, agreed by delegates from more than 160 countries in Marrakech, Morocco.

The agreement, reached at a meeting that began on 29 October, was thrashed out in the early hours of 10 November, after a final 18-hour negotiating session. It has triggered optimism that the treaty will come into force before the World Summit on Sustainable Development, to be held next September in Johannesburg, South Africa.

Contentious questions in Marrakech included the eligibility of industrialized countries to buy and sell rights to emit greenhouse gases, methodologies for reporting and monitoring emissions and emission-reduction activities, and penalties for countries failing to achieve their emission targets.

To come into force, the Kyoto Protocol must be ratified by a minimum of 55 countries that together accounted for at least 55% of global carbon-dioxide emissions in 1990. The United States, the world's largest emitter,

still says that it will not ratify the protocol.

This means that the European Union (EU) and the 'Umbrella Group' of industrialized countries — including Japan, Canada, Australia and Russia — are needed to reach the 55% emissions target. Although the EU is enthusiastic, members of the Umbrella Group have, until now, been reluctant.

To make ratification palatable to Russia, delegates agreed that credits for its natural carbon 'sinks', mainly forests, could be almost doubled to 33 million tonnes of carbon per year. Russia hopes to profit from unused rights to emit greenhouse gases, and now says it will ratify the protocol.

Japan also won concessions that will make market-based mechanisms, such as emissions trading, fully available to all industrialized countries, regardless of whether or not they meet their targets and report their emissions adequately.

But a decision on the legal status of penalties for countries that fail to meet their targets was deferred to the first meeting of parties after the treaty has come into force. ■

Blow to German stem-cell prospects

Alison Abbott, Munich

A cross-party parliamentary commission on bioethics has cast a majority vote against research involving any human embryonic stem (ES) cells in Germany.

The 12 November vote, which came out 17 to 7, will be taken into consideration next January, when the German parliament debates whether to close a legal loophole allowing human ES cells to be imported. In Germany, isolating the cells from human embryos is forbidden.

The situation puts more pressure on the DFG, Germany's main research granting agency, which has twice delayed a decision to release money for a research project on imported human ES cells. The agency

approved the project in principle in May (see *Nature* 411, 875; 2001).

The DFG had hoped that the parliament would have debated the issue before the agency's main committee meeting on 7 December, at which the project was scheduled to be discussed. But events in the United States and Afghanistan have forced ES cells down the political agenda.

Scientists are disappointed by the commission's vote. But they anticipate that the National Ethics Council, which was created by Chancellor Gerhard Schröder earlier this year, will be more favourably disposed towards regulated imports of human ES cells. The National Ethics Council will report next week. ■

Proposal for vaccine lab gets a shot in the arm

Jonathan Knight, San Francisco

The idea of a national vaccine laboratory is gaining momentum in the United States, as advocates of the project pour scorn on the country's ability to produce vaccines. They claim that the recent anthrax attacks have revealed glaring weaknesses in the present system, which effectively relies on drug companies to develop improved vaccines.

But proposals for the new laboratory remain vague. Some suggest that an existing plan for a military vaccine laboratory should be expanded to take a broader role in public health. Another approach would see the Department of Health and Human Services establish a laboratory under the auspices of the National Institutes of Health.

On 5 November, the Institute of Medicine (IOM), which represents leading US physicians and medical researchers, issued a statement calling for a new National Vaccine Authority (NVA). This would develop and stockpile vaccines to meet both military and public-health needs, the IOM suggests. But the IOM has not suggested what it might cost to operate the NVA.

One component of the NVA would be a vaccine laboratory owned by the government but operated by a private contractor, similar in style to the US Department of Energy's existing national laboratory system. Such a facility is described in the 2002 defence authorization bill, which is still being considered by the Congress. But this is only for military use.

The IOM instead advocates building a facility for both public and military purposes. "We think it would be a serious mistake if

it were not a dual-use facility for both the military and the public," says the institute's president, Kenneth Shine.

The lab, as Shine envisions it, would develop vaccines that the private sector has little incentive to investigate. The current vaccine for anthrax, for example, requires six injections over 18 months. But without a guaranteed market, pharmaceutical companies might not make the investment needed to research and test a new one, says Shine.

The proposed laboratory would develop and test new vaccines at public expense, and then offer them to drugs companies for manufacture. And for vaccines that are often in short supply, such as that for tetanus, the NVA would consider options such as stockpiling.

Bruce Gellin, director of the National Network For Immunization Information, a Virginia-based group that advocates public vaccination, says a national authority might have the foresight to prevent shortages. In September, stocks of a newly approved vaccine against *Streptococcus pneumoniae*, which causes respiratory infections in small children that can lead to pneumonia, ran out after the government recommended its use in all children under two years old. "You need to have a place where you could have rational planning and procurement," says Gellin.

But it is the recent fears of bioterrorism that have given new urgency to calls for more government involvement in vaccine development. The crisis has highlighted shortcomings, for example, in US readiness to immunize against smallpox, should this prove necessary. The government last year



On target: the proposed centre would aim to prevent public shortages of crucial vaccines.

commissioned Acambis, a British pharmaceutical company, to produce 40 million doses of smallpox vaccine over 20 years, and is expected to announce a contract this week for the production of 250 million more.

Shine argues that a vaccine authority could maintain a stockpile to deal with various scenarios. "Given military and terrorism needs, it is likely that something in this area is going to happen," he says.

Other groups have put forward proposals similar to the IOM's. A commission chaired by Republican governor James Gilmore of Virginia, for example, recommended building a similar lab in a 31 October report on responses to terrorist attacks.

www.iom.edu

AP/WILL KINCAID

French Nobel protest makes chemist a *cause célèbre*

Sally Goodman, Paris

France's science minister has written to the director of the Nobel Foundation, Michael Sohlman, protesting that French chemist

Henri Kagan of Paris-Sud University was not given a share of this year's chemistry prize.

The unusual move by the minister, Roger-Gérard Schwartzberg, comes after several French scientists voiced their dismay that Kagan missed out.

The coveted prize — which can be shared by no more than three scientists — went to K. Barry Sharpless of the Scripps Research Institute in La Jolla, California, Ryoji Noyori of Nagoya University, Japan, and William Knowles, formerly of Monsanto in St Louis, Missouri, for work on catalytic asymmetric synthesis (see *Nature* 413, 661; 2001).

The French minister's 5 November letter uses arguments taken from an article by Didier Astruc, an organic chemist at the University of Bordeaux, in *Le Monde* newspaper. The letter argues that Henri Kagan was the true pioneer in the field of catalytic asymmetric synthesis, citing his 1971

demonstration of the catalytic asymmetric hydrogenation of olefins in which he produced a large excess of one chiral product.

However, the citation issued by the Nobel Foundation in support of the prize names William Knowles as the founding father of catalytic asymmetric hydrogenation, having first demonstrated the process in 1968. But it states that the chiral product he produced at that time "was modest and hardly of any practical use".

In an interview in the newspaper *Libération*, Léon Ghosez, a chemist from the Catholic University of Leuven in Belgium, denounced the French minister's initiative as "grotesque", adding that the letter would "neither serve the interests of Kagan nor of French chemistry".

Kagan told *Nature* that he is not behind the science minister's initiative and prefers not to comment on the matter.



Chinese plan pins big hopes on small science

David Cyranoski, Beijing

When the Chinese Academy of Sciences presented a self-cleaning 'nano-necktie' to China's president, Jiang Zemin, last December, it provided a boost to the country's high hopes for nanotechnology.

The technology behind the tie, developed by Lei Jiang at the Chinese Academy of Sciences' Institute of Chemistry in Beijing, combines water-repellent and oil-repellent nanometre-scale structures on its surface. Such combinations have a wide range of analogous applications, claims Jiang. And it is this kind of potential that has caught the attention of the Chinese government.

Taking its cue from recently established programmes in the United States and Japan, the government now plans to spend an estimated 2.5 billion renminbi (US\$300 million) over five years on nanoscience.

The scheme includes construction of a National Nanoscience Center in Beijing, scheduled to be fully operational in 2003. Its planners hope that the centre will be a focal point for China's research efforts in nanotechnology, encompassing basic research in nanoscale physics, chemistry and biology.

"It will be a platform that scientists throughout China and the world can use," says Chunli Bai, vice-president of the Chinese Academy of Sciences and a member of the country's national steering commission for nanotechnology.

The centre will also help to bring scientific talent back home to China and foster a multidisciplinary approach to research. "I have watched too many excellent students go overseas, now I want to attract some of them to stay," says Xing Zhu, a physicist at Peking University, who is helping to plan the facility.

"We will have an international advisory committee to establish a world-class laboratory, and we hope to draw many foreign researchers here," Bai adds.

Teams selected by the centre's committee of scientists and government officials will be invited to use the facility, carrying out research on the manipulation of materials at the molecular level.

Officials will next month set the budget for the centre. Construction and equipment are expected to cost between 250 million and 500 million renminbi, and the centre will cost up to 15 million renminbi a year to run. The money will come from several government agencies.

The centre was proposed in a July report by various government bodies, which also called for two separate nanoengineering facilities, in Beijing and Shanghai, for application and commercialization. Plans to establish these centres are in the works.

Technology transfer to industry, either through the nanoengineering centres or

directly, will be a major goal of the nanoscience centre, according to Bai.

"Integrating engineering with science will be challenging in China," says Zhong Lin Wang, a materials researcher at the Georgia Institute of Technology in Atlanta, who has been collaborating with some Chinese universities on nanotechnology. "The key to success

will be collaboration, which has been a weakness in the Chinese community, but I do see some promising changes in recent years."

But some scientists are worried that the nano-necktie has raised expectations for eye-grabbing commercial products. "We should be appealing to real basic science, not this kind of propaganda," one sceptic argues. ■

Polar bears fuel row over Alaskan oil



Out in the cold: field research disputes claims that bears would be harmed by oil drilling.

Mark Schroppe

Polar bears are the latest flashpoint in the intensifying fight over whether the United States should drill for oil in Alaska's Arctic National Wildlife Refuge (ANWR).

President George W. Bush has reaffirmed the case for the drilling — in the name of energy security — since the events of 11 September. But environmentalists, facing an uphill struggle to block it, are threatening to sue the administration for allegedly covering up information on the impact of such drilling on the polar bear population.

Research does not seem to back the environmentalists' claims, however. Recent studies — albeit part-funded by the oil industry — suggest that mother polar bears will not abandon their young as a result of oil exploration and extraction.

Steven Amstrup, a biologist with the US Geological Survey in Anchorage, has surveyed polar bears in Alaska for over a decade. He estimates that about 25 dens are set up by the bears each year in the region of the ANWR where drilling is set to take place.

Environmentalists argue that drilling might disturb some female polar bears so much that they abandon their dens and cubs.

But Amstrup, whose assessment is partly based on data showing that dens are widely dispersed, says that activity in a given area would disrupt few, if any, bears.

In earlier work, Amstrup found that the bears are generally tolerant of outside disturbance (see *Arctic* 46, 246–250; 1993). High-impact activities could also be prevented during the denning season. "There are a lot of options for management," he says.

The Center for Biological Diversity, an environmental pressure group based in Tucson, Arizona, has nonetheless petitioned the US Department of the Interior to release a 1995 assessment of whether the United States complies with a 1973 international agreement to protect polar bears. Congress requested the report but has yet to receive it, as the interior department says it wants to retain it until a new polar bear treaty, agreed with Russia last year, is reviewed by the Bush administration.

Kieran Suckling, the centre's executive director, calls this a "stall tactic" and says the group may use legal means to force the document's release. He believes the report suggests that development in the ANWR could break the 1973 agreement. ■

JEFF FOOT/BBC WILD

Russia establishes advisory council in bid to boost science

Bryon MacWilliams, Moscow

President Vladimir Putin has set up a new advisory council to help him revive Russia's beleaguered science and high-technology sectors.

Scientists hope the body will make science a higher political priority, although they note that it includes many officials who have overseen the free fall in funding and prestige that has made scientists some of the poorest-paid workers in the country.

The Council on Science and High Technology, which comprises 24 politicians and high-ranking science officials, is charged with helping to prioritize policies on science and innovation, as well as drafting legislation. It will meet at least twice a year, and report on developments worldwide in science and technology, promote cooperation with foreign scientific organizations, and recommend measures to ensure the welfare of Russian scientists.

In part, the council represents a reaction to complaints that the Putin administration has failed to address the dire circumstances facing Russia's scientists. "Many have criticized us for the insufficient attention we have paid," says Aleksei Gromov, a government spokesman.

Members of the council, headed by Kremlin official Sergei Abramov, will include Yuri Osipov, president of the Russian Academy of Sciences, Victor Sadovnichii, rector of Moscow State University, and Evgeny Velikhov, president of the Kurchatov Institute in Moscow. ■

Academy backing bolsters shaky support for crust study

Rex Dalton, Boston

EarthScope, a major US initiative to study the Earth's crust that has struggled to win government support, has received a strong endorsement from the US National Academy of Sciences.

"EarthScope will have a substantial impact on earth science in America and worldwide," says an academy report released on 8 November. The report was produced by a seven-member panel chaired by George Hornberger, a hydrologist at the University of Virginia in Charlottesville.

Advocates of the \$400-million project hope the report will nudge the National Science Foundation (NSF) and NASA into incorporating the project in their respective budgets for 2003, which are being prepared and will be released by President George W. Bush in February.

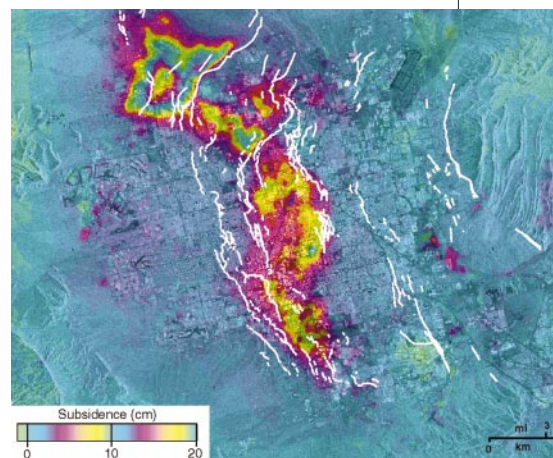
Last year, Congress prevented the Clinton administration from starting EarthScope, and the Bush administration has so far shown little sign of supporting it—although mid-level agency officials are in favour of the plan. "We are doing our best to produce a budget including EarthScope," says Herman Zimmerman, director of the NSF's earth-sciences division.

The NSF component of EarthScope would comprise three main elements: a mobile, national grid of seismometers, called USArray; the Plate Boundary Observatory, which would monitor the movement of tectonic plates in the Pacific Northwest; and the San Andreas Fault Observatory at

Depth (SAFOD), which would drill into the fault south of San Francisco (see *Nature* **405**, 390–392; 2000). These would cost a total of about \$200 million over five years. NASA would contribute a further \$200 million to EarthScope, in the form of an Interferometric Synthetic Aperture Radar (InSAR) satellite, which would monitor land movement.

The American Association of State Geologists, which met in Boston last week, has also endorsed the project, as has the National Science Board, the governing body of the NSF. But supporters of the project remain worried that it may not obtain sufficient political backing. ■

► www.earthscope.org



Rock star: EarthScope would combine satellite imaging with ground-based studies.

Harvard squeaks through oncomouse patent appeal



Mouse protest: activist groups opposed the patent on grounds of animal suffering.

Alison Abbott, Munich

The European Patent Office (EPO) has upheld the patent on the Harvard oncomouse, a mouse strain genetically engineered to be susceptible to cancer.

But the patent office has restricted the breadth of the Harvard University patent—which currently covers all animals that are genetically engineered using the oncomouse technology—to cover only rodents.

More than 100 organizations and individuals had registered their opposition to the patent, which sparked intense controversy when it came into force in 1992.

"The appeal board felt that it was impossible to assume that the balance between benefit to society and suffering to the mouse could be automatically extended to all types of animals," says

Christian Guggerell, the EPO's spokesman on biopatents.

Opponents of the patent consider this to be a tacit admission of unacceptable animal suffering. "The EPO's position lacks logic, both ethically and legally," says Greenpeace spokesman Christoph Then.

The decision is the first to be made by the EPO on an appeal against an animal patent since it adopted the European Union's 1998 directive on biopatents. A moratorium on patents on life had been in effect for two years before that, as the EPO awaited ethical guidelines from the European Union.

The directive allows such patents to be awarded if the benefit to society is deemed greater than the suffering to the animals. Since the directive was adopted, the EPO has granted 20 further patents on animals. ■

Animal-rights activist dies after campaign of hunger strikes

London Militant campaigners against animal experimentation in Britain have gained a martyr to their cause in the form of Barry Horne, who died on 5 November after embarking on his fourth prison hunger strike. Horne, who was jailed for 18 years in 1997 for firebombing shops selling animal furs, had refused food and drink since 21 October.

In 1999, Horne won publicity during a hunger strike held in protest at the government's failure to establish a royal commission to review policy on animal experimentation. At the time, a shadowy group called the Animal Rights Militia



Martyr to the cause? Supporters idolize convicted firebomber Barry Horne after his death.

threatened to kill several prominent British scientists if Horne was allowed to die.

Some representatives of the animal-rights movement warn that Horne's death will harden the resolve of those campaigning against experiments on animals. But a spokeswoman for the Research Defence Society, which backs the regulated use of animals in research, says that there is little sign so far of an increase in activity against individuals or laboratories.

US science funds boosted amid 'stiff competition'

Washington In a victory for advocates of increased spending on non-biomedical science, the US Congress has approved a \$372-million, 8.4% increase in spending on the National Science Foundation (NSF). Its total budget for next year is almost \$4.8 billion, compared to the \$4.47 billion that had been requested by President George W. Bush. Given the stiff competition for money on Capitol Hill, say congressional staff, the NSF allocation is as generous as could be expected.

The spending plan includes \$199 million for nanotechnology, \$75 million for genomic research on economically significant crops, \$241 million for projects in information technology and \$139 million for facilities and large items of research equipment.

Stem-cell registry opens its portal

San Francisco The US National Institutes of Health (NIH) has posted on the Internet its long-awaited registry of human embryonic stem-cell lines, opening the door for grant applications to use the cells. The website, which went live on 7 November, lists 11 companies or research institutions with a total of around 50 cell lines — the exact number remains in question because not all the lines shown on the main page have reference numbers for use in grant requests.

Technical descriptions are available for 17 of the lines; only 5 are ready for distribution. In addition, two providers — Geron of Menlo Park, California, and the Wisconsin Alumni Research Foundation in Madison — list the same lines. The two are embroiled in a legal battle over the rights to distribute them.

To compensate for the delay in posting the registry, the NIH has extended its grant-application deadline until 27 November.

escr.nih.gov

Spanish science official targeted by terrorist bomb

Barcelona The Basque-separatist terrorist group ETA tried to kill a senior official at the Spanish science ministry with a car bomb in

Madrid on 5 November. Juan Junquera — the intended target of the bomb that injured 87 people — escaped with minor injuries.

Junquera is the general secretary of science policy at the ministry. He is responsible for bodies such as CSIC, the Spanish national research council, and occupied important posts in several ministries before the creation of the science ministry last year.

ETA has targeted other senior government officials, although Junquera is the first target associated with research. Two suspects are in custody, after a passer-by saw them running from the scene and informed the police.

NASA given go-ahead for far-sighted mission

Washington NASA's on-off plan to send a probe to Pluto is back on. The US Congress has announced funding for a mission to the last unexplored planet in the Solar System. NASA's 2002 budget includes \$30 million for the Pluto–Kuiper Belt mission, which was not included in President Bush's funding request earlier this year.

Space scientists have campaigned for the project's restoration because there is a limited launch window if the spacecraft is to get a vital gravity boost from Jupiter. This would allow it to arrive before Pluto's elliptical orbit takes it so far from the Sun

that its atmosphere freezes. "This is a victory for public interest," says Louis Friedman, executive director of the Planetary Society, based in Pasadena, California.

Anthrax drugs are the latest target for idle computers

London British chemists plan to enlist computer users across the world to search for better anthrax treatments. Researchers at Oxford University's Centre for Computational Drug Discovery hope to use the 'distributed computing' model, in which spare processing power on idle desktop computers is used to analyse data. They will look for molecules that can inhibit the toxin secreted by the anthrax bacterium, the structure of which was published in last week's *Nature* (414, 229–233; 2001).

The anthrax project builds on an existing initiative run by the Oxford centre, which has already screened some 3.5 billion candidate molecules for anti-cancer properties.

Asteroid danger is not so earth-shattering

Washington In an increasingly uncertain world, here's some reassuring news: after surveying the number of asteroids big enough to destroy civilization if one were to

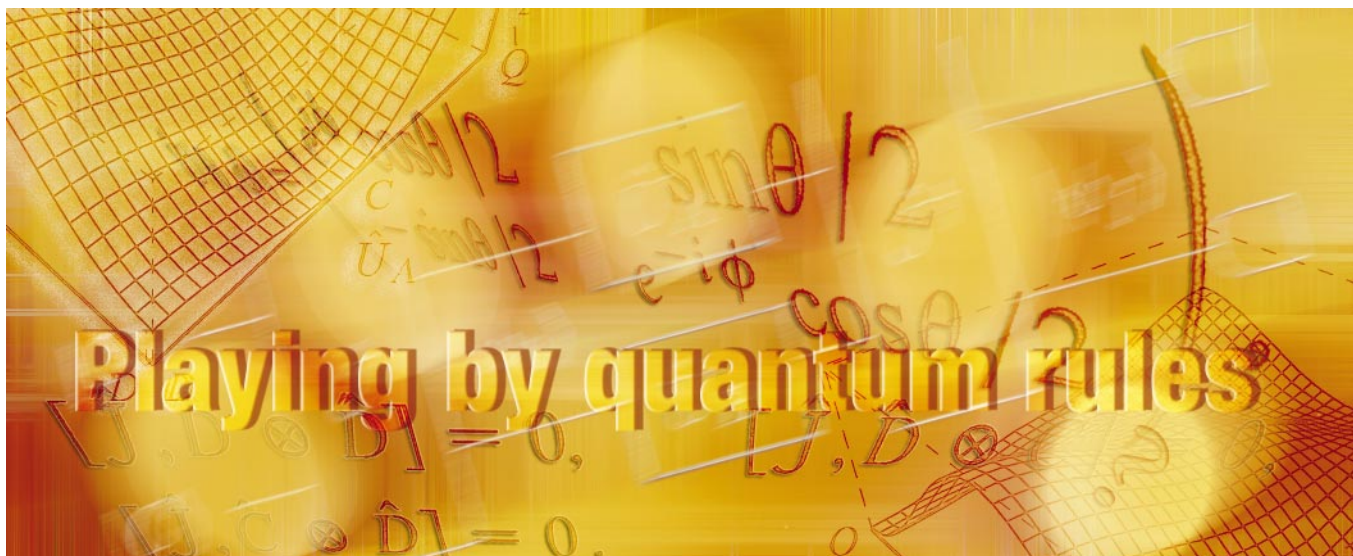


Don't put your house on it: the odds of another huge impact are longer than previous estimates.

collide with Earth, astronomers at Princeton University estimate that the odds of such a collision occurring in the next century have lengthened from 1 in 1,500 to 1 in 5,000.

The team used data from the Sloan Digital Sky Survey to predict that there are some 700,000 bodies over 1 km across in the asteroid belt. They say that previous studies measured asteroid sizes less accurately, leading to an exaggerated estimate. "We should feel somewhat safer," says team leader Željko Ivezić. The results appear in *The Astronomical Journal* (122, 2749–2784; 2001).

► www.journals.uchicago.edu/AJ/journal/contents/v122n5.html



Game theory has been used to study problems from nuclear warfare to animal behaviour. Now physicists are extending it into the quantum realm, opening a new range of potential applications. Erica Klarreich reports.

In the classic puzzle of game theory, two accomplices are awaiting trial, having been caught robbing a bank. Six-month prison sentences lie ahead if both deny involvement. If each blames the other, the police will have enough evidence to send both down for five years. But, crucially, if one keeps silent and the other spills the beans, the informer will get off scot-free while their partner lands a 10-year stretch.

Known as the Prisoners' Dilemma, this puzzle encourages players to adopt destructive tactics. Assuming that the partners are guided by purely selfish motives, the rational option for both is to defect. Both accomplices are then left with five-year sentences, when, had they been able to trust one another and cooperate, they would have got just six months.

Games like these have fascinated mathematicians since the early twentieth century, and they have been used to study a range of subjects from the evolution of animal behaviour to the likelihood of nuclear conflict. Now the field is taking a completely new direction, led by physicists interested in quantum mechanics.

In quantum versions of existing games, new strategies become available to the players. Some researchers say these make the dilemma in the prisoners' game disappear. Others are using the games to develop new algorithms for the quantum computers of the future. In the shorter term, the games could find applications in the stock market.

The first and simplest analysis of a quantum game was published in 1999 by David Meyer, a mathematician at the University of California, San Diego¹. Meyer considered a game in which two players, Q and P, control a

coin that neither can see. Q makes the first and last move. Both players know that the coin starts heads up. On each go, the players can choose to turn the coin over or leave it alone, but neither can see what the other is doing. The game ends after a previously agreed number of moves — if heads shows, Q wins, otherwise P wins. Each player may as well act randomly as there is no other strategy that will improve their chances.

Heading for victory

But transferring the game to the quantum world introduces new possibilities. According to quantum mechanics, objects can exist in mixed states known as superpositions. A hypothetical quantum coin, for example, can exist as heads, tails or a superposition of both. Meyer analysed a new version of the game in which Q is allowed to include quantum combinations of heads and tails,

but P is restricted to classical strategies. This advantage, he found, allows Q to win every time.

Again the coin starts heads-side up, and Q's first move is to put the coin into a state that is an equal mixture of heads and tails. P can then either turn the coin or leave it alone. But whatever P chooses to do, it has no effect on the coin. Q cannot see the coin, but knows that P has failed to alter the coin's state. Q ends the game by making a move that is the exact opposite of the first move, which changes the state from an equal mixture of heads and tails back into pure heads.

Meyer's example might seem abstract, but similar games could help researchers in their quest to find uses for future quantum computers. In 1994, physicist Peter Shor of Bell Laboratories in Murray Hill, New Jersey, electrified the world of computer science by devising an algorithm that would allow a



In a spin: David Meyer's quantum strategy means that a player will always win a coin-flipping game.

M. QUONG

quantum computer — if one is ever built — to factor numbers much more efficiently than any classical computer².

Shor's discovery prompted a search for other quantum algorithms. "People know efficient quantum algorithms are out there," says Neil Johnson, who works on quantum computation at the University of Oxford. But despite a flurry of activity, only a handful of quantum algorithms that can solve problems more efficiently than their classical rivals have been uncovered.

It is here that quantum games might help. Some mathematical problems can be rephrased as games. The problem of searching for a particular entry in a database, for example, can be turned into a game in which players compete to see who finds the entry first. Algorithms for solving the problem then become strategies for playing the game. Finding a quantum algorithm that solves the problem faster than its classical counterpart is equivalent to identifying a quantum strategy that can beat the classical alternative. "This will probably be one of the first benefits of quantum game theory," predicts Johnson. Indeed, Meyer has already succeeded in casting some known quantum algorithms in terms of quantum games³, and is now working on generating new ones.

Cryptography may also benefit. Researchers have built prototype quantum 'cryptosystems' that are, in principle, impossible to crack without the infiltrator being detected. Measuring a quantum system changes its state, so in theory no eavesdropper could examine a message without leaving a trace. But in practice, imperfections in the transmission system introduce noise into the signal, potentially obscuring the presence of an eavesdropper. Quantum game theory might help researchers to understand how vulnerable a system is. "Analysing the security of quantum communication is like analysing a quantum game," says Meyer. "The two people trying to communicate securely are playing against an eavesdropper."

Other researchers are interested in

expanding the scope of the games, and are studying situations in which both players can use quantum strategies. In 1999, for instance, Jens Eisert, then at the University of Potsdam in Germany, and his colleagues described a quantum version of the Prisoners' Dilemma.

In Eisert's scheme, players can choose a superposition of cooperation and defection. The two players' choices are also 'entangled'. According to quantum mechanics, entangled particles — such as two infrared photons formed by the spontaneous splitting of a higher energy ultraviolet photon — exert an influence over each other. If, for example, the polarization of one of the photons is measured, the polarization of the other will be modified instantaneously. Players' choices can be represented by the state of a particle such as a photon or, as in proposed quantum computers, an electron. If the particles are entangled, the two choices can modify one another when measured, even though neither player has knowledge of the other's decision.

Under the influence

This considerably extends the range of possible tactics. Players no longer have to choose straight cooperation or defection, and they can also try to influence the other's choice by means of entanglement. By considering a limited number of all the possible quantum strategies, Eisert showed that rather than straight defection, a form of cooperation in which entanglement allows the players to influence each other's decision was the best option⁴.

But some theorists felt that limiting the range of possible strategies was unnatural. Simon Benjamin of the University of Oxford and Patrick Hayden of the California Institute of Technology have since shown that if two players are allowed the full range of quantum strategies, the best plan is for them to make random choices⁵. However, their analysis of a three-player version of the Prisoners' Dilemma, in which players also

have access to the full range of strategies, shows that a form of cooperation emerges as the best approach⁶.

Such work has stimulated a spate of research on other games, but researchers warn that the subject is still in its infancy. "At this point every new game has to be analysed from scratch," says Meyer. Applications for the work are also some way off, but physicists are pondering potential uses nonetheless.

One example involves 'minority games' — in which players choose between two options and win a pay-off if they are in the minority — which could benefit from being run as quantum games. Trading shares is a example: players benefit if they are one of just a few to invest in a successful stock. Benjamin and Hayden have shown that some minority games yield higher pay-offs overall if all the players are allowed to use quantum strategies⁶.

Building the quantum devices needed to play such games may not be that difficult. Implementing something like Shor's algorithm would require hundreds or thousands of entangled particles, which is technically very challenging. But in simple games such as the quantum Prisoners' Dilemma, only one particle per player is needed. "The technology to play such a game is either present now or is just around the corner," says Seth Lloyd of the Massachusetts Institute of Technology, a physicist who specializes in quantum computing.

At this stage, no one is sure which applications will prove most fruitful. But as the players in this new field hone their skills, the game is on.

Erica Klarreich is an intern with Nature.

1. Meyer, D. *Phys. Rev. Lett.* **82**, 1052–1055 (1999).
2. Shor, P. in *Proc. 35th Symp. Foundations of Computer Science* (ed. Goldwasser, S.) 124–134 (IEEE Computer Society Press, Washington, 1994).
3. Meyer, D. in *Contemporary Mathematics: Quantum Computation and Quantum Information Science* (American Mathematical Society, Providence, Rhode Island, in the press).
4. Eisert, J., Wilkens, M. & Lewenstein, M. *Phys. Rev. Lett.* **83**, 3077–3080 (1999).
5. Benjamin, S. C. & Hayden, P. M. *Phys. Rev. Lett.* **87**, 069801 (2001).
6. Benjamin, S. C. & Hayden, P. M. *Phys. Rev. A* **64**, 030301 (2001).



Winning strategy: Simon Benjamin's multiplayer games could improve the success of share trading.



Cooperative: Jens Eisert has come up with a quantum solution to the Prisoners' Dilemma.

If they could talk the animals...

Too many conservation projects are failing because of ignorance about the behaviour of endangered species. This is why the natural world needs ethologists, says Jonathan Knight.

To the uninitiated, wildlife conservation and animal behaviour seem like two sides of the same coin. Conditioned by television wildlife documentaries featuring singing whales and leaping antelope, we imagine that conservation biologists are well versed in ethology, the scientific study of animal behaviour.

In reality, conservation project teams rarely include a behavioural expert. And ethologists admit that they are partly to blame. "We missed the bandwagon," says Richard Buchholz of the University of Mississippi in Jackson. "It's our own fault."

At the Animal Behavior Society's annual meeting in July, held in Corvallis, Oregon, Buchholz organized a workshop to address this failing. Participants pointed to numerous conservation projects that have run into trouble for lack of behavioural data — but noted a few resounding successes from the marriage of conservation and ethology.

Paul Sherman, a behavioural ecologist at Cornell University in Ithaca, New York, is one of the few ethologists who has helped to redesign a species recovery plan. In 1985, he teamed up with his former student Brad Semel to study a phenomenon in wood ducks (*Aix sponsa*) that conservation managers were calling dump-nesting.

Since the 1930s, managers had been erecting nest boxes on wildlife preserves for wood ducks, which had been hunted to near extinction by the beginning of the twentieth century. Because the birds prefer to nest in tree cavities, it was thought that placing wooden boxes on poles over open marshes would accelerate their recovery.

But many females laid their eggs in the same nest box while other boxes went unclaimed. "We had eggs laid on top of boxes,



Top flight: Paul Sherman's studies of wood ducks helped to stop their population declining.

in the water, even on top of other ducks," says Sherman. Some nests were filled with up to 30 eggs, rather than the natural dozen or so. Because females tend to abandon nests that contain too many eggs, hatching success was low, dipping to only 10% at some sites.

Lame ducks

Dump-nesting turned out to be related to a natural phenomenon known as brood parasitism. Female wood ducks sometimes leave eggs in other birds' nests to avoid the costs of incubating them. Under normal circumstances, a nest might contain two or three extra eggs deposited in this way.

Through many years of painstaking field observations, Sherman and Semel showed that older ducks usually return to the same nest site year after year. But they found that newcomers, particularly yearlings, seem to rely on seeing adult females entering their nest to find a place to lay their eggs. For them, brood parasitism is the norm. At natural nest sites, which are usually hidden in the woods, this does not place too great a burden on the

experienced ducks. But Sherman and Semel found that the nest boxes were just too conspicuous: every yearling in the neighbourhood flocked to them¹.

In response to these findings, conservationists now hide nest boxes in the woods and keep them far apart. Hatching rates among ducks nesting at these sites have soared to more than 70%, much the same as for natural nests².

Unfortunately, says Tim Caro, a behavioural ecologist at the University of California, Davis, the wood duck example is a rare success story. Many ethologists mention the implications of their work for conservation at the end of their papers, but few actually liaise with conservation biologists. "I don't think that it filters through to the decisions being made on the ground," says Caro.

The fact that Sherman and Semel's findings took more than a decade to crystallize illustrates one of the main problems. Conservation projects are often pressed for time and money, says Jim Strittholt, who directs the Conservation Biology Institute, a

consultancy based in Corvallis. "Behavioural studies usually take several seasons," he says. "If you have a nine-month project there is no way you can incorporate that."

Politics can also come into play. Some areas of conservation are so politicized that ethologists have a hard time being heard. In North America, for example, environmentalists, anglers, hydroelectric companies and native Americans all fight to influence decisions over salmon conservation.

Yet this example illustrates the vital contribution that ethologists can make. Hatcheries are now reintroducing endangered salmon species to bolster wild populations. But there is evidence that adult salmon raised in captivity have behavioural abnormalities that could affect their reproduction in the wild. They might spawn too early³, or too far downstream in the river⁴, or compete less effectively for mates than their wild cousins⁵. And if enough captive-bred fish enter a wild population, the reproductive success of the whole population can drop⁶.

Fishy business

Other behavioural scientists are concerned about the emphasis on genetic diversity for salmon reintroductions. The general assumption is that greater diversity makes a population more resilient. But Mart Gross, a behavioural biologist at the University of Toronto, is not convinced.

Females of many species choose males on the basis of gene quality, notes Gross. The result is not maximum diversity, but selection of the best genes. Gross argues that hatcheries might produce more robust fish by letting females choose their mates, rather than pairing the fish off according to a diversity protocol. "Our view is that they actually mongrelize their populations," he says.

Although influencing salmon conservation may be challenging, ethologists who talk to policy-makers are finding they can make a difference. William Sutherland, a biologist at the University of East Anglia in Norwich, is one. Recently passed legislation to open the British countryside to the public has raised concerns about the impact on wildlife, so the government is working to make exceptions for sensitive areas. Sutherland uses theoretical models based on the nesting behaviour of birds to predict the effects of human disturbance⁷. The government has asked him to help to identify areas of farmland and open heath in which populations might be particularly vulnerable. The key is to get involved early on, says Sutherland. "If you come in at the end and say 'Why don't you do this?', it can be seen as a criticism."

This is also the approach of Daniel Blumstein, a behavioural biologist at the University of California, Los Angeles, and the Marsupial Cooperative Research Centre (MCRC) in Sydney, who is collaborating with the South Australian government on



Foxy moves: Andrea Griffin and Chris Evans teach tammar wallabies how to escape predators.

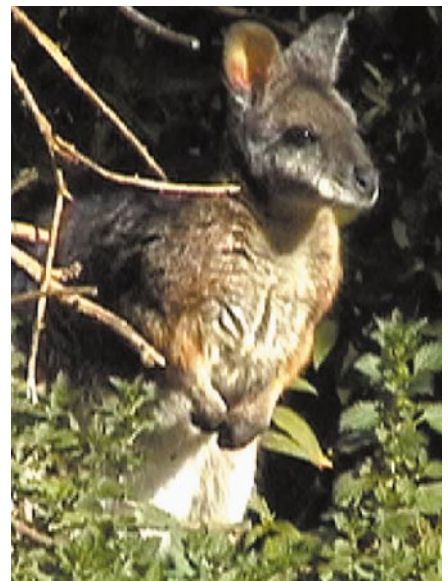
the reintroduction of tammar wallabies (*Macropus eugenii*). Although tammars are no longer found on the Australian mainland, an introduced population survives in New Zealand — where it has no known predators.

This is a potential problem, as animals raised in the absence of predators may struggle to survive when exposed to nature red in tooth and claw. Many reintroduction programmes have failed because of this⁸. Although predator-avoidance behaviour is in part hard-wired, practice makes perfect. Releasing naïve animals into an environment full of predators can be tantamount to putting them out to die, says Blumstein.

Getting jumpy

Through the MCRC, Blumstein is collaborating with Andrea Griffin and Chris Evans of Sydney's Macquarie University to train wallabies to avoid predators. In a typical session they place a stuffed fox in the wallaby's enclosure, then five seconds later a researcher tries to grab the animal — something wallabies particularly dislike. The researchers have found that the training instils a fear of foxes, which the tammar can generalize to other predators, such as cats⁹.

Whether predator-avoidance training improves an animal's chances of survival has been tested in only a handful of species, with mixed results. Houbara bustards (*Chlamydotis undulata macqueenii*)¹⁰ and masked bobwhite quails (*Colinus virginianus ridgwayi*)¹¹ survived longer after training. But Benjamin Beck of the National Zoological Park in



Washington found that it made no difference for golden lion tamarins (*Leontopithecus rosalia*), an endangered Brazilian primate¹².

These variable results underline the need to include a behavioural component in conservation projects. And for ethologists, there is an element of self-preservation, as well as species preservation. "Our lives as scientists depend on the availability of organisms in their natural habitats," says Sherman. ■

Jonathan Knight writes for Nature from San Francisco.

1. Eadie, J., Sherman, P. W. & Semel, B. in *Behavioral Ecology and Conservation Biology* (ed. Caro, T.) 306–340 (Oxford Univ. Press, Oxford, 1998).
2. Sherman, P. W. in *Model Systems in Behavioral Ecology* (ed. Dugatkin, L. A.) 311–337 (Princeton Univ., Princeton, 2001).
3. Nickleson, T. E., Solazzi, M. F. & Johnson, S. L. *Can. J. Fish. Aquat. Sci.* **43**, 2443–2449 (1986).
4. Jonsson, N., Hansen, L. P. & Jonsson, B. *Anim. Behav.* **48**, 740–742 (1994).
5. Fleming, I. A. & Gross, M. R. *Ecol. Appl.* **3**, 230–245 (1993).
6. Fleming, I. A. & Petersson, E. *Nordic J. Freshw. Res.* **75**, 71–98 (2001).
7. Gill, J. A., Norris, K. & Sutherland, W. J. *J. Appl. Ecol.* **38**, 846–856 (2001).
8. Griffin, A., Blumstein, D. T. & Evans, C. S. *Conserv. Biol.* **14**, 1317–1326 (2000).
9. Griffin, A., Evans, C. S. & Blumstein, D. T. *Anim. Behav.* **62**, 577–589 (2001).
10. Van Heezik, Y. & Maloney, R. *Re-introduction News* **13**, 3–4 (1997).
11. Ellis, D. H., Dobrott, S. J. & Goodwin, J. G. in *Endangered Birds: Management Techniques for Preserving Threatened Species* (ed. Temple, S. A.) 345–354 (Univ. Wisconsin, Madison, 1977).
12. Beck, B. B., Castro, M. I., Stoinski, T. S. & Ballou, J. in *The Lion Tamarins: Twenty-five Years of Research and Conservation* (eds. Kleiman, D. G. & Rylands, A.) (Smithsonian Institution Press, Washington, in the press).



Scaled up: greater genetic diversity isn't necessarily better in fish such as these sockeye salmon.

PETER CUPITT

DAN BLUMSTEIN

JEFF FOOTIT/BBC NATURAL HISTORY UNIT

Science is universal, not part of any religion

Islam fostered its rise and can be proud that it has now grown beyond any one culture.

Sir—In your Opinion article “Fighting against terrorism, engaging with Islamic science” (*Nature* **413**, 235; 2001) you tell readers that, as scientists, we can do more for the societies engaged in the present “conflict” (a noble euphemism for an unprovoked attack on civilians), and that the indiscriminate killing of innocents is not the consequence of a clash of civilizations.

You state: “many Islamic scholars and leaders have emphasized that the murder of the innocent is as offensive to their beliefs as to anyone else’s”. Surely the evil of killing people is self-evident and hardly requires a restatement by scholars and leaders, whether Islamic or Icelandic? You go on to say: “resurgent Islam appears to be giving a sense of values ... to populations ... repressed within their countries” and that revolutionary intent (or what others would call ‘terrorism’) is the pastime of only some activist groups. In plain English, you are telling us that one should have understanding for those whose cultural identity is enhanced by flying passenger aircraft into skyscrapers.

Your article provides URLs for two websites as part of your analysis of the conflict between Islamic and Western science. The first contains an original and unbiased discussion of the issue by Dr Mehrzad Boroujerdi; the second is a poorly reproduced speech, based on lecture notes, markedly political and rhetorical in character, but hardly a scholarly document.

Most educated people know that Islamic science is responsible for one of the major events in the history of civilization: the passing on of the Greek science of antiquity to mediaeval Europe. By doing this, Islamic science was not merely a messenger but an active generator of a scientific revolution without which Europe might never have entered the Renaissance. From algebra to algorithm, Islamic science left an indelible print on the Western scientific edifice. This was supplemented by its tolerance for and nurturing of non-Islamic cultures, as demonstrated by the unprecedented flowering of Hebrew scholarship and poetry in Arabic Spain.

But what is meant by Islamic science in the year 2001? Is there Islamic chemistry or Islamic quantum mechanics? I do not think so. I admire and respect an Arab biologist who does superb work in my field, and a geneticist colleague of mine collaborates with Palestinian physicians on a health-related project, even while the guns are roaring. I admire the Arab fellow-scientist because he is first-class, not because he is a representative of Islamic science—a

classification I am certain he would abhor.

Your preaching that we should prevent the killing of innocents by “engaging with Islamic science” is a crass example of neo-colonialist paternalism and an offence to scientists labelled as Islamic. My ancestors made a modest contribution to the compilation of the Bible, but this does not mean that I live by the principle of an eye for an eye; neither did my British Christian colleagues offer the other cheek when threatened by Nazi rule.

The proposal that fostering collaboration between Islamic and Western

Don't use terrorism as an excuse for militarism

Sir—In your Opinion article “Anti-terrorist agendas” (*Nature* **413**, 655; 2001) you seem to have swallowed whole the assumptions about a new ‘war’ promoted by the Bush administration. You discuss how the science community may assist its governments in what I consider a false war.

Your article states: “In times of war, modern governments require a special relationship with their best scientists and engineers—a relationship which now needs to be recast.”

But to have a war there has to be an adversary much more specific and less elusive than the terrorists whom the administration is condemning. In my view the so-called “war” is really an open-ended mobilization for the self-interests of the Bush administration: increased support for the military, increased presence in south-west Asia, decreased civil rights, decreased social spending, muzzling of opposition and blanketing of information.

It is unfortunate that, to promote science, *Nature* believes that we must accede to the politics of an uncivil society. The charge that science and militarism go hand in hand seems here to be borne out. This will only harm science in the long run.

Morton K. Brussel

Loomis Laboratory of Physics, University of Illinois at Urbana-Champaign, 1110 West Green Street, Urbana, Illinois 61801, USA

BSE record set straight

Sir—It was disappointing that your reporter contributed further confusion to the issue of cow brains being tested instead of sheep brains for signs of BSE (“Brain

scientists is an effective means in the fight against terrorism should be first and foremost rejected by those who have been so carelessly categorized as “Islamic” scientists. An international gang of well-financed criminals, using the cream of Western science for the sole purpose of indiscriminate killing, has as little to do with Islamic science as the Boston Strangler has with America’s founding fathers.

Edgar Pick

Julius Friedrich Cohnheim—Minerva Centre for Phagocyte Research, Sackler School of Medicine, Tel Aviv University, Tel Aviv 69978, Israel

mix-up leaves BSE research in turmoil”, *Nature* **413**, 760; 2001) by failing to contact the Laboratory of the Government Chemist (LGC) before declaring that our findings in this case were open to dispute.

To set the record straight, LGC is an independent, non-Government laboratory committed to the highest standards of analytical quality.

We have developed, validated and deployed DNA-based tests for the speciation of meat for many years, often in cases of commercial misdescription of processed meat. The Department for the Environment, Food and Rural Affairs (DEFRA) commissioned LGC to check the species identity of the brain samples that were used in the BSE experiments conducted by the Institute of Animal Health (IAH).

Our tests used two different extraction methods, three methods for qualitative identification and a full battery of controls, blanks and authentic samples.

We were aware that the IAH samples had already been subjected to a number of cycles of freeze-thawing, so the experiments were specifically designed to be able to detect degraded DNA (about 220 base pairs). The identification of beef DNA was unambiguous. (Our report to DEFRA is available on the LGC website, www.lgc.co.uk.)

The assertion in *Nature* that the storage conditions of the samples in transit to LGC could have affected the results does not bear scientific examination. DNA is a robust molecule, and it would be a remarkable sequence of freezing and thawing that turned sheep DNA into beef.

Richard Worswick

Laboratory of the Government Chemist, Queens Road, Teddington, Middlesex TW11 0LY, UK

An unfinished quest

Quirks of the AIDS virus, as much as human bungling, are stalling progress.

Shots in the Dark: The Wayward Search for an AIDS Vaccine

by Jon Cohen

W. W. Norton: 2001. 440 pp. \$27.95

Dennis R. Burton

With 22 million dead from AIDS and the threat of countless millions more to follow (current infection rates are estimated at 100,000 people per week), the search for a vaccine against the human immunodeficiency virus (HIV) is one of the most urgent and important priorities facing biomedicine. This quest has been followed by the science writer Jon Cohen since its beginnings in the early 1980s, and is now documented in *Shots in the Dark*. The title and its subheading — “The wayward search for an AIDS vaccine” — give notice that this is a critical account. Jon Cohen has been reporting on AIDS for the journal *Science* for some ten years and is very highly regarded by researchers in the field.

His book is a long one, totalling more than 400 pages, and is woven from two main strands. The first is a detailed account of the principal events and characters of AIDS-vaccine science and politics, from the first brash announcement in 1984 that there would be a vaccine within three years, to the present day. The second strand is the author's own search for why we still lack a vaccine, where he considers we have gone wrong, and how we could do better. The historical account is fascinating, very well researched and sourced, and should be required reading for anyone wishing to understand the efforts to develop an AIDS vaccine. From White House press secretary Larry Speakes' joking about gay cruising and AIDS in 1982, through to the present controversy about testing subunit vaccines in the United States and Thailand, Cohen presents an authoritative, highly readable and objective narrative.

The author's search within a search is provocative, but in my opinion somewhat less persuasive. Let me not overreach here. Cohen struggles constantly to deal fairly with difficult issues and never resorts to cheap shots or preaching. He focuses on two themes. One is the suggestion that AIDS-vaccine research has lacked central direction (for which the National Institutes of Health (NIH) is criticized). Cohen calls for a ‘Manhattan’ project, analogous to the concentrated effort of scientists from many subdisciplines to develop the atomic bomb. The second theme is the struggle between ‘empiricists’, who favour early and many clinical trials, and ‘reductionists’, who desire more information before committing to large-scale trials.

The call for a Manhattan project has some



The virus's formidable arsenal of evasive strategies is a major challenge for AIDS-vaccine workers, seen here examining a plate used to measure immune cells.

merit and there have been increasing moves in that direction, particularly from the International AIDS Vaccine Initiative. One needs, however, to temper all criticisms of the NIH, given that this organization is one of the great success stories of the twentieth century. AIDS research in all areas has been built on the rock of NIH funding, and most HIV vaccine research has been, and will continue to be, funded by the NIH. Certainly, the organization suffers from the inflexibility inherent in a large federal bureaucracy and mistakes have been made, but it is not responsible for a ‘wayward search’ for an AIDS vaccine. The author sometimes sees the problems of AIDS-vaccine development too much in terms of human bungling and not enough in terms of the quirky properties of the virus itself, which has evolved a formidable arsenal of strategies for avoiding immune responses — thereby presenting major problems for vaccine developers.

The empiricism versus reductionism argument is a complex one, sometimes unhelpfully simplified by key protagonists in the book. The late Jonas Salk, for instance, lamented scientists who “mainly seek answers to the most basic questions — dissecting the squirrel to see why it climbs the tree — and there is no detectable push to translate those data into experimental vaccines”. This notion of scientists in ivory towers is more and more a myth. If altruism does not provide the necessary push, then

surely, in the modern age of biotechnology, the profit motive will.

Others accuse the NIH culture of undervaluing applied research, but then many scientists would argue that there must be something worthwhile to apply. In AIDS-vaccine research, until recently, there have been very few data of sufficient promise to support a vaccine efficacy trial. In 1995, a meeting was organized at which the assembled scientists “all understood that Thailand urgently needed to try something — even something that had only an outside chance of working”. Is it wise to spend millions of dollars on clinical trials that have only an outside chance of working?

To return to empiricism versus reductionism, all science has a strong empirical component. Every experiment is something of a shot in the dark: if we knew what the result of an experiment was going to be, we wouldn't need to do it. Nevertheless, we constantly make judgements about the worth of following a particular path or taking a particular shot in the dark. Not all shots can or should be taken. It would seem that Cohen joins the contingent who believe that judgement in the HIV-vaccine field has been made too often in favour of restraint. Some of us disagree — arguing that at least the gun should be loaded with live ammunition before shooting in the dark — but doubtless the controversies will continue.

As a scientist, I had a number of other

different perspectives from the author. Sometimes I wanted to plead with him that it is not only the guy frenetically running around organizing, planning and talking about vaccine programmes, but also the bench scientist working to solve some crucial problem, who is making big contributions. I was irritated at times by his tendency, Hollywood-style, to embrace mavericks, some of whose work has not stood the test of time. I was also puzzled by the omission of any reference to the solution of the structure of the virus surface protein gp120, which I believe will eventually be seen as a major milestone on the route to an HIV vaccine. But, reservations aside, my emphatic overall conclusion is that this is an excellent book which demands to be read. ■

Dennis R. Burton is in the Department of Immunology, Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, California 92037, USA.

Fungal fables

Slayers, Saviors, Servants, and Sex: An Exposé of Kingdom Fungi
by David Moore

Springer: 2001. 175 pp. \$29.95, £20

Nicholas J. Talbot

Collecting wild mushrooms is not generally viewed as a pastime for the truly adventurous. But there is a group of thrill-seeking mushroom collectors who attempt to distinguish between two almost identical species: one, *Amanita caesarea*, is perfectly edible, whereas the other, the death cap mushroom *Amanita phalloides*, will invariably kill you. Indeed, eating just one cubic centimetre of the death cap is thought sufficient to kill the average person. This story of the lunatic fringe of mycologists (presumably a rapidly dwindling group) is one of the anecdotes recounted in David Moore's new celebration of the fungi. The author is a big fan of moulds in all their various forms, and conveys his enthusiasm consistently throughout this quirky but entertaining book.

Fungi are important. This is the central message of Moore's book and one that he illustrates with numerous examples. Fungal diseases kill humans, kill crops and rot harvested foods. In describing the spectacular epidemics and diseases caused by fungi the book is compelling. The symptoms of St Anthony's fire, for example, are given in grisly detail: burning fevers, hallucinations, swollen limbs, severe inflammation followed by numbing of the extremities and complete loss of fingers, toes, arms and legs. All are caused by the ergot fungus *Claviceps purpurea*, consumed in contaminated rye

The injured and the ingenious

A case of brittleheart has decimated the sycamore tree (*Platanus occidentalis*) shown on the right. Its trunk has been hollowed out by the action of fungi, putting it at risk of collapsing with the next bout of windy weather. The strange-looking tree shown below is the aptly named sack-of-potatoes or Socotran desert rose tree (*Adenium obesum sokotranum*), which is found on the Indian Ocean island of Socotra. Its swollen trunk stores essential water. *Trees* by Roland Ennos (Natural History Museum, London, £9.95, \$14.95) explores the world of this most familiar group of living organisms.



bread. One outbreak of the disease reputedly killed 40,000 people in France in AD 944, and ergotism has even been suggested to have been the cause of symptoms that took hold of eight young girls in Salem and led to the witch trials of 1692. This story, and the production of potent toxins such as α -amanitin and aflatoxin, are used to demonstrate that fungi demand our respect.

The beneficial applications of fungi, from gastronomy to bioremediation, are given equal coverage. Fungi are, of course, well known as producers of antibiotics, and the story of the discovery of penicillin has been told many times. But our modern dependence on antibiotics is poignantly illustrated here by the story of Albert Alexander, a British policeman who died in 1941 after being scratched by a rose thorn. Initial treatment with the newly available penicillin made him well enough to sit up in bed, but he quickly succumbed to his infected wound once supplies of the drug ran out. The discovery and application of the immunosuppressant cyclosporin, and the best-selling cholesterol-lowering agents, the statins — both natural fungal products — are similarly well described.

Although relatively light-hearted and anecdotal in style, there is a serious underlying theme throughout the book, because David Moore is clearly not a happy man. He

is unhappy that fungi are viewed only as curiosities and consequently occupy, in his view, the peripheries of mainstream science. Ignored by biology departments pensioning off mycologists, by funding agencies, simplistic politicians and international development bodies, fungi offer tremendous untapped potential but are constantly overlooked. Having devoted a career to their study, Moore wants us to know he is annoyed. He also describes the history of mycology as littered with the 'nearly' men of science: the Reverend Miles J. Berkeley, who studied the origin of the potato-blight epidemic and anticipated the germ theory of Louis Pasteur by almost 25 years, but did not prove his case; and Agostino Bassi who, similarly, recognized fungi as agents of disease at an even earlier date.

It is here that I take issue with Moore's thesis. As this year's Nobel prize for medicine aptly demonstrated, fungi are at the centre of scientific endeavour. From the one-gene-one-enzyme theory to the control of cell division, we have learned a huge amount of fundamental biology from fungi. Indeed, this is the most striking omission in the book, because if there is one thing about fungi of which the public are really uninformed, it is their use in biomedical research. Fungi, both yeasts and filamentous species, are superlative experimental

organisms, and their widespread use in biological research means there are more mycologists working today than ever before.

I would recommend this book particularly to those who know, or currently care, little about fungi. They will learn a lot in this easy read, and they can even try the anagrams at the back of the book. Here I learned that 'fantastic ugly mess' is an anagram of 'fungal systematics', which should upset the taxonomists. ■

Nicholas J. Talbot is at the School of Biological Sciences, Washington Singer Laboratories, University of Exeter, Exeter EX4 4QG, UK.

Elements of galactic evolution

The Chemical Evolution of the Galaxy

by Francesca Matteucci

Kluwer: 2001. 308 pp. £72, \$106

Stuart Ross Taylor

Walter Baade, at Palomar during the Second World War, observed that the halo of the great spiral Andromeda galaxy was dominated by red, metal-depleted stars, whereas its disk contained mainly blue, metal-enriched stars. (Astronomers, to the exasperation of chemists, persist in using 'metals' to describe all the elements heavier than helium. But the term, along with 'metallicity', is too thoroughly entrenched to be replaced). Baade's observations enabled Olin Eggen, Donald Lynden-Bell and Allan Sandage to erect their classical model of galactic evolution, in which a spherical halo collapses to a disk. But it was Beatrice Tinsley who, in the course of her brief life, focused attention on the chemical evolution of galaxies, and stimulated the explosive growth in this field.

All of this was possible because of the elegant explanation of how chemical elements are synthesized — surely one of science's greater achievements — by Margaret and Geoffrey Burbidge, William Fowler, Fred Hoyle and Alastair Cameron. This, in turn, was based on an understanding of the relative abundances of the chemical elements, first established by Victor Goldschmidt from meteorite data.

Now the topic of galactic chemical evolution has become a major field of research. Francesca Matteucci's book is the second to address the subject, following on from the definitive 1997 text by Bernard Pagel, *Nucleosynthesis and Chemical Evolution of Galaxies* (Cambridge University Press). David Arnett's monumental work, *Supernovae and Nucleosynthesis* (Princeton University Press, 1996), is weighted more towards nucleosynthesis — the cosmic formation of atoms through nuclear reactions

in the cores of stars and through supernova explosions. However, it deals specifically with the chemical evolution of galaxies in a final chapter.

Matteucci's book usefully complements these two, more advanced texts, concentrating on the Milky Way. She begins with an assessment of the observational evidence for chemical evolution, now so greatly enhanced by the development of high-resolution spectroscopy, and continues with a major section on stellar evolution and nucleosynthesis. As a planetary geochemist, I found her clear and precise treatment very useful. Matteucci concludes with a major discussion of the formation and evolution of the Milky Way, finishing with a brief assessment of the age of the Galaxy and some comparisons with other spiral galaxies.

As is typical in the astronomical literature, the text is replete with acronyms. Readers unfamiliar with the subject will be relieved to discover that, in the relatively peaceful field of astrophysics, IRA refers to 'instantaneous recycling approximation' and ICM to 'intracluster medium', not to their deadly alternatives. And they will discover that IMF, the 'initial mass function', has nothing to do with finance. These caveats aside, the book is an example of straightforward writing that might well be emulated by native writers in English. There is no index, but this surprising omission is mostly covered by a very detailed table of contents, so that it is easy to find one's way around.

The development of the field in the past two decades is impressive, Matteucci herself having made significant contributions. The earlier concepts have been replaced by a model of the Milky Way in which the Galaxy is divided into four distinct stellar populations, respectively the halo, bulge, thick disk and thin disk, which do not have simple relationships with one another. Thus we learn that the "evolution of the thin disk is completely disentangled from the evolution of the inner halo".

So galaxies have a much more complex history than previously imagined. Like terrestrial continents, they have undergone collisions and have grown by accretion of disparate material. Our Milky Way Galaxy appears to have evolved along at least two independent paths. One was the classical collapse from a spherical halo to a disk. And superimposed on this was accretion of gas and other galaxies, so that the present Galaxy contains components that formed in several distinct environments. As the accreting gas may be deficient in heavy elements, younger stars forming from this late-arriving material will not necessarily show the expected enrichment in metals that would be predicted to occur had the Galaxy evolved in isolation.

In an interesting assessment of the age of the Galaxy, Matteucci concludes that 13–15

New in paperback

The Eureka Effect: The Art and Logic of Breakthrough Thinking

by David Perkins

W. W. Norton, \$14.95, £11.95

Wilderness and the American Mind

by Roderick Frazier Nash

Yale University Press, \$16.95, £11.99

The New World of Mr Tompkins

by George Gamow & Russell Stannard

Cambridge University Press, £10.95, \$16.95

"Although presenting quantum mechanics accurately to a broad lay readership will certainly remain a challenge for some time, *The New World of Mr Tompkins* is a lovely book which I am sure everyone interested in modern physics, from the age of 11 upwards, will enjoy enormously." Anton Zeilinger, *Nature* **405**, 618 (2000)

Newton's Tyranny: The Suppressed Scientific Discoveries of Stephen Gray and John Flamsteed

by David Clark & Stephen P. H. Clark

W. H. Freeman, \$14

billion years seems probable, with the age of the disk being in the range of 8–10 billion years. There is a useful discussion of nucleochronology — a technique for dating stages in stellar nucleosynthesis by examining the relative abundances of radioactive isotopes. Also discussed is the promise shown by measurements of thorium/europium ratios. It seems likely that these will yield ages for stars which are based on the radioactive decay of thorium-232, and will be totally independent of ages based on the Hubble expansion rate of the Universe — that is, once the formidable technical problems have been overcome and we have a "specific and correct model for the growth and chemical evolution of the solar neighbourhood".

Although Matteucci notes the interesting fact that the Sun is more metal rich than most similar stars in the solar neighbourhood, she does not comment on the observation of Guillermo Gonzalez that the planets so far discovered beyond the Solar System have formed around the more metal-rich stars. This also raises other interesting questions. As the metallicity of stars varies with location in the Galaxy, so the regions in which habitable planets may form may likewise be limited. The most favourable location (in which we find ourselves) is apparently the thin galactic disk, as discussed by Peter Ward and Don Brownlee in *Rare Earth* (Copernicus, 2000). ■

Stuart Ross Taylor is in the Department of Geology, Australian National University, Canberra, ACT 0200, Australia.

The world of testable truths

On Science

by Brian Ridley

Routledge: 2001. 240 pp. £30, \$50 (hbk);

£7.99, \$12.95 (pbk)

John Casti

What is scientific knowledge? In what way does it differ from any other type of knowledge? Are there logical limits to the kind of knowledge we can acquire about the world around us by employing the methods and procedures of science? For that matter, what is science itself? These are the types of questions that physicist Brian Ridley addresses in this thought-provoking and well-written volume.

The book opens by reminding us of the crucial fact that there are many types of truths, of which the testable truths of science stand beside the revealed truths of religion, the persuasive truths of the humanities and the demonstrable truths of mathematics. Ridley then asserts that there is yet another kind of truth, which he defines as "magical truth". This is a kind of truth that is complementary to that of science — one

associated with the non-material, human forces in the world. Poetry, music and the fine arts are examples of where these magical truths reside. Such truths are the domain of the magus, and not the scientist, who is focused on the recurrent in the world, not the unique. This juxtaposition of science and magic is a theme woven throughout the book.

Taking up E. O. Wilson's view of science as an enterprise that gathers knowledge and condenses it into testable laws and principles, the author raises several objections to the idea of a scientific 'Theory of Everything' that could — even in principle — answer all the questions we might pose about the world. First, there is the limiting fact that quantum theory, our best scientific theory thus far, involves the inherent uncertainty associated with any measurement of a physical system. Next comes the self-referential fact that the very tools we use to probe nature are themselves part of nature. And finally there is the most serious limitation of all, science's inherent inability to cope with anything unique — sometimes labelled 'origins problems'. As science is in the business of addressing recurrent phenomena, it's a debatable matter as to whether questions about the origin of the Universe, life, language, humanity or anything else should even

be considered a question within the purview of science.

The realm of mathematical truths is the one indisputable place where we can probably pose questions that cannot be answered using the normal modes of mathematical argumentation. The Gödel Incompleteness Theorem ensures that this is the case: it states that in any consistent logical system, there are statements that can be made which can be neither proved nor disproved within the framework of that system. The linkage between this realm and that of the physical world of quarks, planets and billiard balls is crucial for understanding the limits to science. Ridley provides an excellent introduction to these matters.

From mathematics to mind is but a small leap in Ridley's odyssey. The bridge is the long-running controversy between philosophers and computer scientists over the possible existence, even in principle, of a computing machine that could think just like you and me.

Gödel's theorem makes its appearance here again as part of the book's account of Roger Penrose's well-known — and perhaps equally well-criticized — arguments against the protagonists of 'strong' artificial intelligence (AI). Ridley clearly favours the anti-AI camp's position that a thinking machine is essentially a contradiction in terms, and that thought, human-style, is just that: human-style. Period.

The final part of the book is in many ways the least interesting, mainly comprising an assortment of fairly obvious observations about the interface of science both with the humanities and with society in general. The fact that science cannot answer questions about ethics, art or religion is hardly news. Neither is the fact that the gap separating literature and science can be bridged only by recognizing the complementary aspects of science and art. So although one can certainly applaud the author for re-emphasizing these well-known facts, they hardly merit much attention.

But this is a minor quibble, and one cannot ask for miracles in a 200-page volume covering such a broad spectrum of themes and issues. On balance, this is an excellent introduction to a number of important questions residing at that interface where philosophy borders on science, and science becomes philosophical. One can only hope that the author may one day write a sequel, in which the issue of the limits to scientific knowledge will be addressed in much more detail, complete with examples of how one might identify those questions that are beyond the bounds of scientific reasoning — and why!

John Casti is at the Institute of Econometrics, Operational Research and System Theory, Technical University of Vienna, Argentinierstrasse 8, A-1040 Vienna, Austria.



Oddities from the cupboard

The collected illustrations of Albertus Seba's *Cabinet of Natural Curiosities* (Taschen, DM300, £100, \$150) were first published in the early eighteenth century. Seba's 'curiosities' comprise the exotic species he bought from sailors and travellers returning to the ports of Holland. It was the heyday of exploration, and the beginning of trade between continents. Taschen has republished the illustrations in a luxurious edition of astonishing beauty. The book also includes essays in English, German and French. (*Cabinet of Curiosities/Das Naturalienkabinett/Le Cabinet des curiosités naturelles d'Albertus Seba.*)

Twice as natural

Richard E. Lenski

According to the book of *Genesis*, God made all life and “created man in his own image”. It is no wonder, then, that humankind has long sought to create life of its own. But what is artificial life?

Setting aside for a moment the meaning of ‘life’ itself, the use of the word ‘artificial’ introduces its own quandary. Where does ‘natural’ end and ‘artificial’ begin? Thomas Browne, in *Religio Medici* (1642), said that “all things are artificial; for Nature is the Art of God”. This view might seem to exclude science. But by emphasizing the artistry of creation, Browne encouraged the study of “Nature, that universal and publick Manuscript, that lies expans’d unto the eyes of all”.

A radically different idea of artificial life appeared a few years later. In *Leviathan* (1651), Thomas Hobbes said that “Nature ... is by the Art of man ... imitated, that it can make an Artificial Animal”. Just what sort of artificial animal did Hobbes have in mind, that humans could have created 350 years ago? In his view, the political state is a kind of super-organism, “an Artificiall Man; though of greater stature and strength than the Naturall, for whose protection and defence it

was intended”. To Hobbes, who sought to avoid anarchy, life in this artificial state is much better than the natural condition, in which life is “solitary, poore, nasty, brutish, and short”.

Hobbes also wondered whether “all Automata (Engines that move themselves by springs and wheeles as doth a watch) have an artificial life”. It now seems quaint to suggest that a watch is alive because it can move itself. But where should we draw the line between animate and inanimate, between life and non-life? Is a robot that walks more alive than a ticking watch? What if the robot can think? What if it has emotions and drives? What if it can reproduce itself? HAL, the spaceship’s computer in Arthur C. Clarke’s novel *2001: A Space Odyssey*, was intelligent and driven to self-preservation. Yet without the capacity to reproduce, would we say that HAL was alive?

The importance of reproduction for understanding life, even artificial life, found its way into William Paley’s *Natural Theology* (1802), in which Paley advanced the creationist argument from design. According to this view, the intricate and interconnected features of organisms that suit them to their environments imply the existence of a creator, in the same way in which the intricately connected parts of a watch imply a watchmaker “who comprehended its construction and designed its use”. Paley took the argument another step by asking the reader to imagine a watch that “possessed the unexpected property of producing, in the course of its movement, another watch like itself”. (He remarked that “this thing is conceivable”, although I doubt that the pun was intended.)

Paley thought that the capacity for self-replication should “increase beyond measure our admiration of the skill which had been employed in the formation of such a machine”. He invoked his hypothetical watch to rebut the counter-argument that, although a watch has a watchmaker, organisms are made by their parents. For Paley, self-replication merely pushed back in time the actions of the creator. Humans have not yet created any entirely self-replicating robots, although one might reasonably view a self-replicating program — a computer virus — as alive in this sense.

Paley imagined descent without modification. Computer viruses, likewise, have no intrinsic capacity to evolve — new varieties are created by a few malicious hackers. Does self-replication without evolution constitute life, as understood by biologists today?

Charles Darwin was the first to grasp firmly the importance of descent with

Artificial life

The meaning of ‘artificial life’ has evolved through the years. We now have artificial life that can itself evolve.

modification. He closed *On the Origin of Species* (1859) by saying that “from so simple a beginning endless forms most beautiful and wonderful have been, and are being, evolved”. We now know that real organisms make heritable errors, or mutations, during self-replication. Many mutations are harmful and are eliminated by natural selection, but some are fortuitously beneficial, thereby providing the raw material for evolutionary adaptation.

It is fortunate that computer viruses, so far, cannot spontaneously evolve. But other self-replicating computer programs are subject to copying errors. And these mutations allow populations of programs to evolve in, and adapt to, their operating environment. The biologist Tom Ray wrote the Tierra software with precisely this evolutionary potential in mind. He found that programs quickly evolved into more efficient replicators, including parasites that replicated themselves by exploiting others. Chris Adami, a physicist, and Charles Ofria, a computer scientist, then developed Avida, in which programs evolve a computational metabolism that is above and beyond their capacity to self-replicate.

Alas, these ‘digital organisms’ are not nearly as sophisticated as simple bacteria, or even real viruses. But then, artificial life has been around for only a few years. The natural stuff has been replicating, mutating and evolving for billions of years. Who knows what forms artificial life will take in a thousand, a million or a billion years? ■

Richard E. Lenski is in the Center for Biological Modeling and the Center for Microbial Ecology, Michigan State University, East Lansing, Michigan 48824, USA.

FURTHER READING

Adami, C. *Introduction to Artificial Life* (Springer, New York, 1998).
Chao, L. *BioScience* **50**, 245–250 (2000).
Ray, T. S. in *Artificial Life* Vol. 2 (eds Langton, C. G., Taylor, C. E., Farmer, J. D. & Rasmussen, S.) 371–408 (Addison-Wesley, Redwood City, California, 1991).
Wilke, C. O., Wang, J., Ofria, C., Lenski, R. E. & Adami, C. *Nature* **412**, 331–333 (2001).

WEBLINKS

Digital Life Laboratory
♦ <http://dillab.caltech.edu/avida>
Evolved virtual creatures
♦ <http://www.genarts.com/karl/evolved-virtual-creatures.html>



Monsters of the imagination: Lewis Carroll’s Hare said that Alice was “twice as natural” as he imagined. Should we be similarly impressed by today’s evolving artificial life-forms?

Self-assembly lights up

John D. Joannopoulos

Opals do it, even biomolecules do it, so why can't self-assembly be harnessed to create photonic crystals with near-perfect order? A new technique shows that absolute order may not require absolute control.

Photonic crystals have emerged as a potentially powerful platform for making light do things that were previously impossible¹. For example, optical light could be guided through air rather than through optical fibres, which would reduce losses. Photonic devices, from high-speed switches to low-power microlasers, would be useful ingredients in fibre-optics communications and in future approaches to high-speed optical computing. In a nutshell, the idea is to design periodic structures that affect the behaviour of photons in much the same way that crystalline semiconductors affect the properties of electrons^{2,3}.

Central to this concept is the formation of a photonic 'band gap' — a range of frequencies for which light is forbidden to exist within the bulk of the photonic crystal. The presence of a band gap depends on a particular periodic structure within the crystal, but whereas the periodic arrangement of atoms occurs naturally in semiconductors, photonic crystals need to be fabricated artificially. On page 289 of this issue, Vlasov *et al.*⁴ exploit an almost natural self-assembly technique to grow a template for making a high-quality photonic crystal. That they can do this in a way that both limits and controls the number of structural defects is truly impressive.

The challenge of fabricating photonic band-gap structures is enormous, as the lattice constant of the photonic crystal (the size of the periodic unit cell) must be comparable to the wavelength of the light passing through the crystal. Optical communications systems, for example, operate in the near-infrared, which means that the photonic crystal-lattice constant must have dimensions of roughly a micrometre. Although this is some 2,000 times larger than the lattice constant of atomic crystals, it is still over 100 times smaller than the average diameter of a human hair.

At these micrometre length scales, controlling the fabrication of intricate structures is far from straightforward. As photonic crystals are usually made from dielectrics, which are insulating materials such as air, one successful approach has been to make them literally out of air by creating an 'inverse' crystal within a matrix of another material. Providing the matrix has a high enough refractive index (silicon is a good example) it will lead to the formation of a photonic band gap. The photonic 'crystal' is

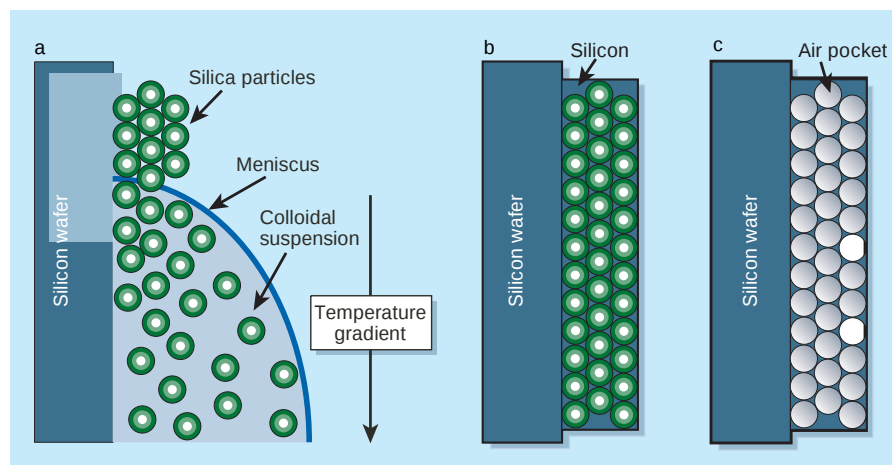


Figure 1 Growth in photonics. Vlasov *et al.*⁴ show how to create a silicon photonic crystal starting from the self-assembly of a colloidal suspension of silica microspheres into an ordered template. **a**, Silica particles are forced into an ordered arrangement on the surface of a vertical silicon wafer as the meniscus is swept downwards by evaporation of the solvent. An applied temperature gradient allows convective flow of the particles that helps minimize unwanted sedimentation. **b**, Once the silica structure, or template, is formed, the voids between the microspheres are uniformly infiltrated with silicon. **c**, Removal of the silica spheres by wet etching leaves behind the desired silicon photonic crystal structure attached to the silicon chip.

the matrix material containing the periodic array of air-filled spaces.

Perhaps even more challenging is the creation of photonic crystals with intentional defects that allow photons to enter the band gap. Why, might you ask, would you want to do this? A defect, or imperfection, in an otherwise perfect periodic structure can allow light to be trapped in the vicinity of the defect while being excluded from the medium surrounding it. So, a point-like defect could serve as a microcavity to localize light, a line-defect as a linear waveguide to guide light, a planar defect as a planar waveguide, and so on⁵. These properties provide a new mechanism for moulding or controlling the properties of light. In this sense, defects are good things in photonic crystals, at least if they are deliberately designed that way: herein lies some of the exciting potential of photonic crystals.

To accomplish these goals, researchers have focused attention essentially on two very different approaches to fabrication: nanolithography and self-assembly. Nanolithography is already widely used by the semiconductor industry and is an expensive, precise and painstaking step-by-step fabrication process that enables the calculated construction of photonic crystal structures, defects and defect networks. Although nanolithography

is highly complex, two recent successes^{6,7} — involving fabrication of photonic crystal structures with sub-micrometre lattice constants — attest to its great potential.

By contrast, the spontaneous crystallization of colloids (hard particles in suspension) provides a much simpler, faster and cheaper alternative to nanolithography^{8–13}. This approach involves the 'natural' self-assembly of polymer or silica microspheres from colloidal suspension into a solid, three-dimensionally periodic structure, which can then act as a template for construction of the photonic crystal. The template contains voids that can be infiltrated by a material of high refractive index, and removal of the original template by wet etching (by dissolving with acid for example) leads to the desired photonic crystal structure. The drawbacks to this approach are the limited number of crystal structures that can be formed, the purity with which they can be formed, and the difficulty of incorporating specifically designed defects. Unfortunately, there is no difficulty whatsoever in incorporating many unwanted defects in these structures, and this has made people sceptical of the practicality of this approach.

The work of Vlasov and collaborators⁴ is an exciting new development in colloidal

self-assembly that enables fabrication of large (roughly 1 cm) single crystals of a given photonic structure, with negligible amounts of unwanted defects. It had previously been shown¹⁴ that crystallization of sub-micrometre particles at a meniscus between a vertical substrate and a colloidal suspension leads to thin, planar, well-ordered structures as the meniscus is slowly swept across the substrate by evaporation. But sedimentation of larger particles away from the meniscus is usually faster than the process of solvent evaporation, so this approach was believed to be limited to particles with diameters less than 0.4 μm . For commercial (near-infrared) applications the particles have to be at least twice as large. Vlasov and collaborators overcame this problem by cleverly adding a temperature gradient (and therefore inducing convection) along the suspension. The convective flow combats sedimentation and provides a continuous flow of particles to the meniscus region (Fig. 1a).

Using this technique, Vlasov *et al.* have succeeded in ordering silica spheres (0.86 μm in diameter) onto a silicon wafer. The resulting template has an order of magnitude fewer point defects and the individual silica crystals are two orders of magnitude larger than previously possible. Once the silica template is in place, the voids in the structure can be infiltrated with a dielectric (Fig. 1b). Silicon was again used here and deposited using a commercial chemical vapour deposition approach. The silica spheres are subsequently removed by wet etching to leave an air-filled silicon lattice (Fig. 1c). The signature of a photonic band gap was confirmed by optical spectroscopy in two directions through the silicon lattice.

But Vlasov and colleagues did not stop here. Having greatly reduced the number of intrinsic defects, they introduced intentional defects by adding a trace amount of silica spheres of a different diameter to the original suspension. Because all the silica spheres are eventually removed after the silicon is in place, this leads to the creation of selected air-filled defects in the final photonic crystal structure. Of course, the positions of these designer defects is random at the moment, but the authors are exploring various ways of providing the necessary spatial control.

Finally, Vlasov *et al.* show that the photonic crystal film can be patterned using standard photolithography and ion-etching techniques. This is important if silicon-based photonic components and existing silicon microelectronics are to be integrated onto a single chip, leading to advanced optoelectronic devices that will speed up the next generation of telecommunications and computers. Although many challenges remain before practical optical components and devices become reality, Vlasov and colleagues' work is an exciting new development in the self-assembly route to microphotronics. ■

John D. Joannopoulos is in the Department of Physics, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, Massachusetts 02139, USA.

e-mail: joannop@mit.edu

1. Joannopoulos, J. D. *et al.* *Nature* **386**, 143–149 (1997).
2. Yablonovitch, E. *Phys. Rev. Lett.* **58**, 2059–2062 (1987).
3. John, S. *Phys. Rev. Lett.* **58**, 2486–2489 (1987).
4. Vlasov, Y. A., Bo, X.-Z., Sturm, J. C. & Norris, D. J. *Nature* **414**, 289–293 (2001).

5. Joannopoulos, J. D. *et al.* *Photonic Crystals* (Princeton Press, New Jersey, 1995).
6. Lin, S. Y. *et al.* *Nature* **394**, 251–253 (1998).
7. Noda, S. *et al.* *Science* **289**, 604–606 (2000).
8. Astratov, V. N. *et al.* *Nuovo Cim. D* **17**, 1349–1354 (1995).
9. Wijnhoven, J. E. & Vos, W. L. *Science* **281**, 802–804 (1998).
10. Vlasov, Y. A. *et al.* *Adv. Mater.* **11**, 165–169 (1999).
11. Braun, P. V. & Wiltzius, P. *Nature* **402**, 603–604 (1999).
12. Müller, M. *et al.* *Adv. Mater.* **12**, 1499–1503 (2000).
13. Blanco, A. *et al.* *Nature* **405**, 437–440 (2000).
14. Jiang, P. *et al.* *Chem. Mater.* **11**, 2132–2140 (1999).

Gene regulation

Code of silence

Judd C. Rice and C. David Allis

Using genetic approaches and a filamentous fungus, molecular biologists have found further evidence of an intimate relationship between genomes and the histone proteins that evolved to package them.

Every cell in your body contains the same genetic material. Yet your cells are not physically identical, in part because they switch different genes on and off according to where the cells are in the body and what information they are receiving. One way of controlling which genes are expressed is by 'epigenetic' mechanisms, which allow identical genetic material to exist in different on or off states that can be transmitted faithfully over many cellular generations, and even from egg and sperm to embryo. Active and silent gene expression states can be defined by certain epigenetic 'marks' within the genome.

One well-studied epigenetic mark associated with gene silencing is DNA methylation — the enzymatic addition of methyl groups to cytosine nucleotides in DNA. Although we now understand how this silences genes, the mechanisms that bring about DNA methylation have remained elusive. But on page 277 of this issue, Tamaru and Selker¹ provide one possible answer. They show that DNA methylation in the fungus *Neurospora crassa* depends on methylation of different molecules — histone proteins².

Cells use histones to compress DNA into a compact structure known as chromatin, which serves to organize and package the genome into the nucleus. The core histone proteins (two each of histones H2A, H2B, H3 and H4), along with the DNA that wraps around them, define the nucleosome — the fundamental unit of chromatin. Of particular interest to Tamaru and Selker's study¹ is histone H3, whose long amino-terminal tail extends well outside the nucleosome and, like other histone tails, is subject to an array of chemical modifications of the amino-acid side chains, including acetylation and methylation. These modifications on histone proteins affect the access of key regulatory factors and complexes to chromatin and thereby influence gene expression.

For example, there are some gene-silencing

processes that require both the deacetylation of histones and DNA methylation³. In this case, methylation by DNA methyltransferases leads to the recruitment of methyl-binding proteins, which in turn associate with other repressive proteins including histone deacetylases (Fig. 1a). The deacetylases remove acetyl groups from lysine amino acids within histone tails — a process that often correlates with silent regions of genomes. That model predicts that DNA methylation influences histone modifications, which somehow results in gene silencing. But Tamaru and Selker¹ show that, at least in filamentous fungi, the opposite is true: histone modifications influence DNA methylation.

Using genetic approaches to study *N. crassa*, Tamaru and Selker discovered a gene, which they named *dim-5*, essential for both DNA methylation and gene silencing. Next, they analysed the putative protein sequence encoded by *dim-5* and identified a region that is very similar to an amino-acid motif known as the SET domain⁴. In some proteins, this domain can methylate a specific lysine residue (found at position 9 in the amino-acid chain) in histone H3. This modification has been shown to be involved in the formation of silent regions of the genome in organisms ranging from yeast to humans^{5,6}. Tamaru and Selker also found that the DIM-5 protein methylates histone H3 in biochemical assays. The findings show that DNA methylation relies on DIM-5 and, by extension, on histone methylation (Fig. 1c).

To validate their findings *in vivo*, the authors inserted genes encoding histone H3 with mutated lysine 9 into wild-type *N. crassa* cells. The cells displayed a consistent loss of DNA methylation and silencing, indicating that lysine 9 is required for DNA methylation. It is also essential to *N. crassa* — those cells that survived the 'transformation' procedure had just one copy of the mutant H3 gene, as well as the original, wild-type copy. In other words, too much of the lysine 9 mutant

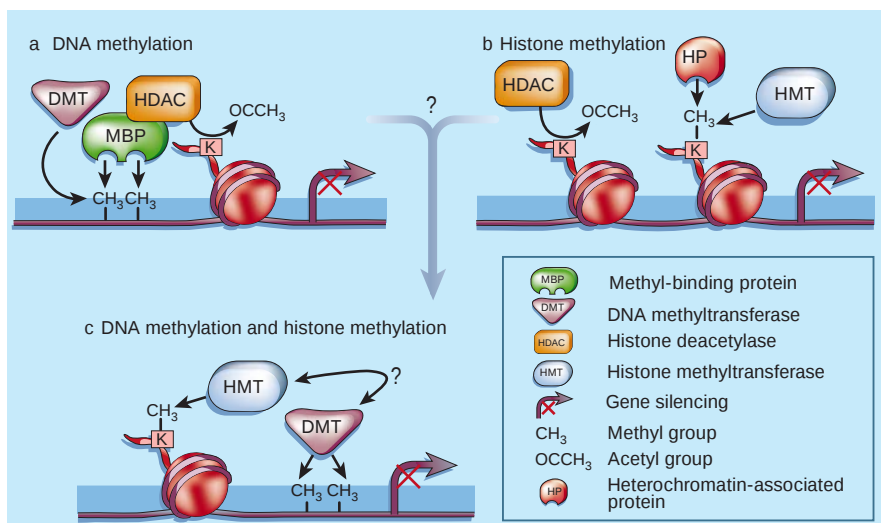


Figure 1 Methylation-associated models of gene silencing. **a**, DNA methylation. Cytosine nucleotides are methylated by a DNA methyltransferase, increasing the affinity of methyl-binding proteins for DNA. Some of these proteins associate with repressive complexes that contain histone deacetylases, resulting in the removal of acetyl groups from lysine (K) residues in the tails of histone proteins, and gene silencing. **b**, Histone methylation. Acetyl groups must be removed before lysine residue 9 on histone H3 can be methylated by a histone methyltransferase. Heterochromatin-associated proteins preferentially bind methylated lysine 9, leading to gene silencing. **c**, DNA methylation and histone methylation. Tamaru and Selker¹ show that methylation of histone H3 lysine 9 is required for subsequent DNA methylation and gene silencing in filamentous fungi. It is not known whether or not the histone and DNA methyltransferases interact.

gene resulted in cell death. Interestingly, though, mutation of this amino acid had no noticeable effect on another type of fungus, *Saccharomyces cerevisiae*⁷, and indeed the methylation of DNA and of histone H3 lysine 9 has not been detected in this species⁸. This implies that the histone- and DNA-methylation-dependent mechanism of gene silencing may not apply to all organisms.

That message was also clear from the discovery that a silencing pathway using histone methylation can function independently of DNA methylation⁶. Fruitflies (*Drosophila melanogaster*) and another yeast, *Schizosaccharomyces pombe*, have no known regions of DNA methylation in their genomes. In these organisms, gene silencing is associated with histone deacetylation and histone methylation followed by the binding of a repressive protein (a 'heterochromatin-associated protein'; Fig. 1b)^{6,9,10}. Such repressive proteins are typically localized to silent regions of the genome¹¹.

By contrast, the DNA-methylation silencing pathway has not yet been found to work independently of histone methylation, supporting Tamaru and Selker's proposal¹ that histone methylation is required for DNA methylation. As yeast and many invertebrates do not typically have DNA methylation, perhaps this phenomenon evolved with histone methylation to add another layer of gene regulation in more complex genomes (Fig. 1c).

We predict that the mechanisms of gene silencing in humans are hybrids of the models illustrated in Fig. 1. But these models may not be mutually exclusive in any species. For

example, DNA methylation is not used to silence many of the genes that are switched on and off in a periodic manner in the cell-division cycle. So, histone methylation may have a predominant or exclusive role in keeping these genes silent until required, as shown for the *cyclin E* gene¹² in human and other cells. In contrast, histone and DNA methylation in combination may be used to repress constitutively silenced genes, as Tamaru and Selker show. In female vertebrates, this type of silencing may also occur on the inactive X chromosome, where the methylation of both histone H3 lysine 9 and DNA is enriched (our unpublished results).

Tamaru and Selker's exciting results¹ show once again how important histone methylation is in the epigenetic control of gene expression. It will now be necessary to re-evaluate fundamental epigenetic concepts with respect to histone methylation, and to unravel exactly how DNA methylation and histone methylation interact. Will the histone and DNA methyltransferases associate in a repressive protein complex *in vivo*? Could methyl groups on histones recruit DNA methyltransferases, or do DNA methyl groups recruit histone methyltransferases? Perhaps more important, it remains to be seen whether this newly discovered mechanism of gene silencing takes place in humans. If so, there will certainly be many ramifications for human development and disease. ■

Judd C. Rice and C. David Allis are in the Department of Biochemistry and Molecular Genetics, University of Virginia Health Sciences Center, 1300 Jefferson Park Avenue, Box 800733,



100 YEARS AGO

Photography for Naturalists. The advantages of photography as compared with wood-engraving for the illustration of works on natural history are in many ways so great that any attempt to perfect and popularise the methods in use should be heartily welcomed. Quite apart from artistic effect, the great superiority of photography is that it ensures absolute accuracy, and, when living animals are the subjects, shows them in natural attitudes. In wood-engraving there are several sources of error which only too frequently make themselves apparent. In the first place, the draughtsman may make a blunder. But too often it is the engraver who is in fault, very frequently from mistaking the nature of some feature in the drawing he has to reproduce. For example, the author of the volume before us calls attention to a curious engraver's error in a well-known popular work, where, from some misconception, the mouth of a stickleback appears in a totally wrong position.

From *Nature* 14 November 1901.

50 YEARS AGO

The progress of nuclear physics in the past two decades has depended greatly on the use of machines for producing high-energy particles; these fast particles are used for bombarding atomic nuclei with the aim of producing nuclear transformations. Though the first machines were built nearly twenty years ago, much still remains to be known about the atom. It must be remembered that we know today about nine hundred different kinds of atomic nuclei; only three hundred of them are found in Nature, while the remainder have to be prepared artificially, using either high-energy particles or neutrons coming from a nuclear pile. Furthermore, each of these nuclei is a complicated structure, with features many of which are still not understood. Finally, quite new phenomena have been recently discovered, such as the artificial creation of mesons in high-energy nuclear collisions. These mesons, previously found only in the cosmic radiation, can now be studied much more closely, and their study is thought to bear closely on the riddle of the 'nuclear forces', the forces whereby the particles in an atomic nucleus are held together... there are at least three kinds of mesons with different mass, each of them unstable and prone to transform itself within a small fraction of a second into two photons.

Otto Frisch

From *Nature* 17 November 1951.

Charlottesville, Virginia 22908-0733, USA.

e-mail: allis@virginia.edu

1. Tamaru, H. & Selker, E. U. *Nature* **414**, 277–283 (2001).
2. Rice, J. C. & Allis, C. D. *Curr. Opin. Cell Biol.* **13**, 263–273 (2001).
3. Ng, H. H. & Bird, A. *Curr. Opin. Genet. Dev.* **9**, 158–163 (1999).
4. Jenuwein, T., Laible, G., Dorn, R. & Reuter, G. *Cell. Mol. Life Sci.* **54**, 80–93 (1998).
5. Rea, S. *et al. Nature* **406**, 593–599 (2000).

6. Nakayama, J., Rice, J. C., Strahl, B. D., Allis, C. D. & Grewal, S. I. *Science* **292**, 110–113 (2001).
7. Mann, R. K. & Grunstein, M. *EMBO J.* **11**, 3297–3306 (1992).
8. Jenuwein, T. & Allis, C. D. *Science* **293**, 1074–1080 (2001).
9. Bannister, A. J. *et al. Nature* **410**, 120–124 (2001).
10. Lachner, M., O'Carroll, D., Rea, S., Mechtler, K. & Jenuwein, T. *Nature* **410**, 116–120 (2001).
11. Eissenberg, J. C. & Elgin, S. C. *Curr. Opin. Genet. Dev.* **10**, 204–210 (2000).
12. Nielsen, S. J. *et al. Nature* **412**, 561–565 (2001).

Condensed-matter physics

Dressing up bare particles

Hartmut Haug

In a plasma, a charged particle attracts a cloud of oppositely charged particles that modifies its behaviour. The first direct observations of this ultrafast 'dressing' process confirm key tenets of quantum kinetic theory.

Particles are rarely free. In condensed matter, such as a metal conductor, electrons cannot drift freely but instead have to adjust to other particles and fields around them — a process called 'dressing'. The particle becomes, as Lev Landau called it, a 'quasi-particle', whose properties may be quite different from those of the original 'bare' particle. Describing how these properties arise from the interaction of particles with their environment has been one of the major tasks of modern condensed-matter theory.

Now, on page 286 of this issue, Huber *et al.*¹ report the first direct experimental observations of the dressing of charged particles on an ultrafast (femtosecond) timescale. At these timescales (a femtosecond is 10^{-15} s, which is to a second what a second is to 32 million years) it is clear that the dressing process is not instantaneous, but takes a short period of time, as suggested by quantum kinetic theory. This technique paves the way to new experiments capable of separating out many-particle interactions that occur on different timescales.

A well-known quasi-particle in solid-state physics is the polaron — an electron surrounded by a virtual cloud of phonons, which are the quanta of the tiny vibrations of the crystal lattice. In essence, a moving electron distorts the crystal lattice, and is in turn influenced by a cloud of lattice phonons. Particularly strong polaron correlations occur in superconductors, in which an electron dressed by lattice phonons affects the behaviour of another electron; the two electrons overcome their mutual repulsion and their collective motion leads to resistance-free flow of current.

In a plasma of charged particles, such as negatively charged electrons and positive ions, the particles are no longer bound together as atoms or molecules, but can interact in new ways. In free space, the strength of interaction between two charged particles is given by the Coulomb potential,

but in a plasma, collective effects come into play and the Coulomb potential is modified by correlated interactions. Each charge carrier attracts a cloud of oppositely charged particles, which work to screen the mutual repulsion between like charges as much as possible, thereby lowering the overall Coulomb interaction. A charged particle surrounded by a cloud of opposite charges is a dressed quasi-particle. Collective oscillations of the plasma caused by the positive and negative quasiparticles vibrating against each other are known as plasmons. Such plasmons can be detected in semiconductors from the way they scatter infrared light.

Huber *et al.*¹ observe the evolution of a plasma of charge carriers in a semiconductor and watch the formation of charge screening and the build-up of plasmons on a femtosecond timescale. To monitor how the many-particle interactions develop in time

they use the technique of ultrafast laser spectroscopy², which involves probing the system with ultrashort laser pulses.

The authors study a dense plasma of negatively charged electrons and positive 'holes' — the charge carriers within the semiconductor gallium arsenide (GaAs). To create the electron-hole plasma they apply an ultrashort (10 fs) laser light pulse to a GaAs crystal (Fig. 1a). This pulse is in the visible range and is absorbed by the crystal. The initially randomly distributed electrons and holes will quickly redistribute to lower the total Coulomb energy. A second laser pulse, lasting 27 fs, is then transmitted through the crystal; this pulse is in the infrared (10^{12} Hz) range and is not absorbed in unexcited GaAs (Fig. 1b). With this pulse the authors are able to measure the 'complex dielectric function' — an indicator of how easily electrons can be separated by an applied electric field — which provides direct information about how far the screening or dressing has developed.

By varying the time between the two laser pulses, the authors obtain a picture of how the dielectric function changes over a very short time interval: a resonant structure (due to emerging plasmons) develops as the delay time between pulses increases from 0 to 175 fs (see Fig. 3 on page 288). Within 70 fs of the original light pulse the particles are fully dressed (Fig. 1c).

What does this experiment tell us? There are several theoretical papers predicting what might happen to non-equilibrium systems, such as an excited plasma, on ultrafast timescales. For example, my theory group^{3,4} has calculated the screened Coulomb potential between two particles in a non-equilibrium system. The observations of Huber *et al.* fully support the theoretical predictions.

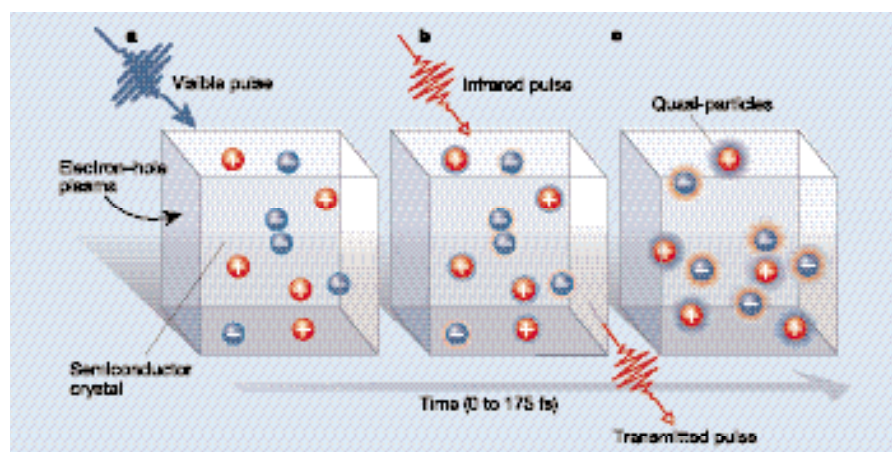


Figure 1 Watching particles get dressed. In the experiment of Huber *et al.*¹, the formation of quasi-particles is observed on a femtosecond timescale. a, An ultrashort laser pulse of visible light is used to excite a plasma of positively charged holes and negatively charged electrons in a semiconductor crystal. b, A second pulse of infrared light is transmitted through the crystal to monitor the build-up of charge screening in the plasma. Charge screening occurs because the electrons and holes modify their mutual interactions in order to lower the Coulomb energy as much as possible. c, Within 70 fs of the original light pulse the screening is completed, as indicated by the oppositely charged clouds around the particles.

These observations also test a central result of quantum kinetics^{5–7} — a theory developed to describe the kinetics of quantum systems on femtosecond timescales. Over such ultrashort time intervals, the behaviour of plasmons and phonons, for example, can no longer be described by semiclassical Boltzmann kinetics, but has to be described by quantum kinetics, which takes the coherent wave nature of the particles into account. This quantum coherence causes memory effects in the scattering integrals that are very different from the simple instantaneous Boltzmann scattering rates. These memory effects are also responsible for the delayed build-up of charge screening observed by Huber *et al.* in their experiment.

It would be of great interest if the charge-screening dynamics in the semiconductor could be measured as the plasma density increases to the point at which the frequencies of the lattice phonon and the collective plasmon oscillations coincide. The resulting state is a mixed phonon–plasmon

mode, whose behaviour would be an entirely new topic in ultrafast physics. The same technique could also be used to study the dynamics of the build-up of magnetic polarons in high-temperature superconductors. There are so many systems involving large numbers of particles (from the vibrational dynamics of large biomolecules to nuclei containing lots of protons and neutrons) that many interesting questions remain unexplored. ■

Hartmut Haug is at the Institut für Theoretische Physik, Johann Wolfgang Goethe-Universität, D-60054 Frankfurt, Germany.
e-mail: haug@itp.uni-frankfurt.de

1. Huber, R. *et al.* *Nature* **414**, 286–289 (2001).
2. Chemla, D. S. & Shah, J. *Nature* **411**, 549–557 (2001).
3. Bányai, L., Vu, Q., Mieck, B. & Haug, H. *Phys. Rev. Lett.* **81**, 882–885 (1998).
4. Vu, Q. & Haug, H. *Phys. Rev. B* **62**, 7179–7185 (2000).
5. Haug, H. & Jauho, A. P. *Quantum Kinetics in Transport and Optics of Semiconductors* (Springer, Berlin, 1996).
6. Kuhn, T. in *Theory of Transport Properties of Semiconductors* (ed. Schöll, E.) (Chapman and Hall, London, 1998).
7. Bonitz, M. (ed.) *Progress in Nonequilibrium Greens Function Theory* (World Scientific, Singapore, 2001).

Evolutionary biology

Counting on immunity

Theo C. M. Bakker and Marc Zbinden

Certain variable immune proteins affect an animal's choice of mate. In some species, females pick males with proteins as dissimilar as possible to their own. Studies of sticklebacks now reveal another mechanism.

In vertebrates, a crucial part of the immune system consists of proteins of the so-called major histocompatibility complex (MHC). There is staggering diversity in the genes that encode these proteins: for example, some MHC genes in human populations come in several hundred different variants (alleles). Yet it is not at all clear how such variability within populations is maintained.

On page 300 of this issue, Reusch and colleagues¹ provide compelling evidence, based on their studies of sticklebacks, for a previously unknown mechanism: females 'count' MHC alleles when looking for a mate. MHC proteins are thought to influence an animal's odour, and, given a choice between the odour of males with many or a few different MHC alleles, female sticklebacks prefer the males with greater numbers. So, by ensuring greater diversity within individual offspring, this 'sexual selection' mechanism also maintains extensive variation within populations.

The MHC genes encode cell-surface proteins — the MHC molecules — that present peptides derived from pathogenic organisms to T cells. This is a crucial step in the immune response to the pathogens. MHC molecules also shape an individual's repertoire of T cells, by displaying peptides from 'self' proteins to immature T cells in the thymus. Any

T cells that react to the self-peptides should die, preventing an autoimmune reaction. MHC proteins are divided into two classes: MHC I molecules present peptides from within an individual's cells (such as viral or self peptides), whereas MHC II molecules display peptides from sources taken up from outside the cell (such as bacterial or self peptides). Typically, each class consists of several duplicated genes, so each individual expresses several different MHC molecules that differ with respect to the peptides they can present. For example, Reusch *et al.*¹ could distinguish up to eight different MHC II alleles in individual stickleback populations.

So, how is this diversity in MHC alleles in populations maintained? Several potential mechanisms have been proposed, involving natural or sexual selection^{2–5}. Pathogens could drive natural selection for MHC variability in two ways. First, an individual with many different MHC alleles can respond to a wider array of pathogens than an individual with less diversity, and so is more likely to survive and pass on its genes to the next generation. Second, because pathogens try to evade the immune system by adapting to common MHC alleles, individuals with new or rare MHC alleles have an advantage.

Sexual selection would maintain MHC variation in a different way. MHC composition somehow influences an individual's odour⁶, so females may secure genetic (including MHC) diversity for her offspring by choosing to mate with males with a certain smell. Reusch *et al.*¹ have revealed a new mechanism that influences a female's choice of mate.

The idea that sexual selection underlies MHC diversity is not new. It dates back to a 1976 paper by Yamazaki *et al.*⁷, who showed that male mice preferred to mate with females with combinations of MHC alleles that were dissimilar to their own. Since then, such 'disassortative' mating has been repeatedly established for mice, and has also been suggested for humans⁵. Reusch *et al.*¹ explicitly tested whether or not female sticklebacks prefer males with dissimilar MHC combinations, but found that the females did not care whether they had many or few MHC alleles in common with the males. Instead, they simply preferred males with many MHC alleles. The authors investigated the mating choices of female sticklebacks that were ready to spawn by presenting the females with water from two tanks, each of which had contained a different male. Females were most attracted by the odour of males with many MHC alleles.

What makes sticklebacks behave differently from mice and humans? So far, no one has tested the MHC-allele-counting option in these species (for example, in previous mouse studies, strains with limited MHC variation were used). But sexual-selection mechanisms for upholding MHC variation may partly depend on social structure. MHC-dissimilar mating in mice is thought to be a mechanism for avoiding inbreeding, because individuals recognize the odours of former nest-mates^{2–6}. This makes sense for populations of house mice, which are structured in family groups. Reproducing sticklebacks are not organized in such groups, and inbreeding is less likely. Moreover, stickleback nests can contain thousands of embryos from different mothers and even fathers. So the familiar odours of former nest-mates are no reliable indication of kinship. The allele-counting mechanism described by Reusch *et al.*¹ is the first known MHC-dependent sexual-selection process that is not confounded by inbreeding avoidance.

MHC diversity is thought to be beneficial, but you can have too much of a good thing. Too much MHC variation in individuals is associated with both an increased risk of autoimmunity and a reduced T-cell repertoire⁵. So you would expect allele-counting mating partners to try to optimize the number of different MHC alleles in their offspring. If so, a female's choice would depend on how many MHC alleles she has herself. In fact, Reusch *et al.*¹ did find that female



Figure 1 Sex and the single stickleback. Male sticklebacks with many different MHC proteins are most likely to be chosen as mates by females.

sticklebacks with many different MHC alleles tended to be less attracted by males with high MHC diversity.

So, to work out the relative contributions of allele counting and disassortative mating in maintaining MHC diversity in different species, population geneticists need to model both scenarios, taking into account social structure and the costs of autoimmunity. MHC-dissimilar matings can explain the degree of MHC variability in mice⁸. But intuitively, you might think that allele counting is a stronger selective force to favour rare alleles and so maintain MHC diversity.

Female choice may affect many different male characteristics⁹; male sticklebacks, for

example, are well known for their visually conspicuous traits (Fig. 1), some of which are associated with parasite resistance¹⁰. Traits such as degree of redness and body size have an obvious influence on female choice. Such traits may suggest whether or not the male will provide good care to the embryos in the nest, or they may have evolved simply for reasons of attractiveness. But visually conspicuous traits might also reveal a male's MHC make-up or non-MHC genetic background, which could in turn affect the male's MHC-dependent ability to resist pathogens². If so, sexual selection would count even more on MHC.

Theo C. M. Bakker and Marc Zbinden are at the Institute for Evolution and Ecology, University of Bonn, An der Immenburg 1, D-53121 Bonn, Germany.

e-mails: tbakker@evolution.uni-bonn.de
mzbinden@evolution.uni-bonn.de

1. Reusch, T. B. H., Häberli, M. A., Aeschlimann, P. B. & Milinski, M. *Nature* **414**, 300–302 (2001).
2. Apanius, V., Penn, D., Slev, P. R., Ruff, L. R. & Potts, W. K. *Crit. Rev. Immunol.* **17**, 179–224 (1997).
3. Potts, W. K. & Wakeland, E. K. *Trends Genet.* **9**, 408–412 (1993).
4. Potts, W. K. & Wakeland, E. K. *Trends Ecol. Evol.* **5**, 181–187 (1990).
5. Penn, D. J. & Potts, W. K. *Am. Nat.* **153**, 145–164 (1999).
6. Penn, D. & Potts, W. K. *Trends Ecol. Evol.* **13**, 391–396 (1998).
7. Yamazaki, K. *et al. J. Exp. Med.* **144**, 1324–1335 (1976).
8. Potts, W. K., Manning, C. J. & Wakeland, E. K. *Nature* **352**, 619–621 (1991).
9. Bakker, T. C. M. & Pomiankowski, A. *J. Evol. Biol.* **8**, 129–171 (1995).
10. Barber, I., Arnott, S. A., Braithwaite, V. A., Andrew, J. & Huntingford, F. A. *Proc. R. Soc. Lond. B* **268**, 71–76 (2001).

Statistical mechanics

A departure from equilibrium

David Ruelle

Non-equilibrium processes, such as heat conduction in a bar of metal, remain poorly characterized at the microscopic level. Detailed analysis of simple models can introduce a new degree of understanding.

What do we mean by equilibrium? Take a bar of metal in equilibrium at room temperature and watch it closely. Nothing happens. Now impose an electric potential difference at the ends of the bar: electric current flows, the bar emits some heat, and it is no longer in equilibrium. If we try to describe the physics of our metallic bar at the microscopic level (the level of atoms and electrons) we find that our fundamental understanding of equilibrium is very good, but our understanding of non-equilibrium states is pretty poor. In one case we can make precise and useful calculations; in the other, hardly any. This situation is unfortunate because the phenomena of life occur far from equilibrium. After 100 years of limited progress in extending the theory of equilibrium to non-equilibrium systems, there are signs that we are finally getting somewhere. In particular, a

paper in *Physical Review Letters* by Bernard Derrida, Joel Lebowitz and Eugene Speer¹ provides a simple model of a non-equilibrium system that permits very detailed calculations. They find that the fluctuations around a non-equilibrium steady state are very different from those at equilibrium.

Let us first review the theory of equilibrium founded by Maxwell, Boltzmann and Gibbs more than 100 years ago. When looking at a system at thermal equilibrium we can calculate the probability of a given state having a particular energy by using the Boltzmann formula, which is based on the absolute temperature of the system, T , and the Boltzmann constant, k . When there are many possible states, the probability of a state having energy close to E is known as $d(E)$, a measure of the density of states around energy E . The formula for $d(E)$ tells

us that the system is at thermal equilibrium (most states have energies near the equilibrium value, E_0) when $d(E_0)$ is at a maximum, or when the free energy of the system is at a minimum. (The free energy is the amount of energy available for work when the system undergoes a change.) The formula also gives the probability that there are deviations or fluctuations from that value. So at equilibrium there are simple formulas to compute the equilibrium value of the energy, and to study fluctuations around this value. At this level of description the size of the system does not matter, but the energy is a sum of local contributions.

Away from equilibrium we have to start from scratch. In their work, Derrida *et al.*¹ discuss a very simple model: a metallic bar consisting of i lattice sites ($i = 1 \dots N$). Each site is either empty or occupied by an electron. Each electron jumps randomly (once per second on average) to the next site on the left or right, provided it is empty. The probability that the left or right end of the bar is occupied is ρ_0 and ρ_1 , respectively. Assuming that $\rho_0 > \rho_1$ (the non-equilibrium situation) there will be a net flow of electrons from left to right. For large N we can introduce a spatial coordinate $x = i/N$, and calculate the 'density profile' of the system, $\rho(x)$, along the metallic bar. When the system is in a steady state (but not under equilibrium conditions) the optimal density profile turns out to be $\rho(x) = \rho_0(1-x) + \rho_1$. This is the macroscopic behaviour that one might expect: a non-equilibrium steady state with an occupation probability that is linearly related to x .

What about the fluctuations? As Derrida *et al.* show, they are totally unlike those at equilibrium. To analyse their model further the authors use a somewhat mysterious matrix method. They find that the optimal density profile $\bar{\rho}(x)$ has fluctuations with probabilities proportional to $\exp\{-NF(\rho)\}$, where F is the non-equilibrium free-energy functional. A functional is a 'function of a function': the free energy is a function of the density profile, which is itself a function of position. This is the point at which the formulae become too complex to reproduce here.

But the expression for $F(\rho)$ is more than a technical *tour de force*. Unlike the free energy in the equilibrium case, $F(\rho)$ is a non-local functional. This means that the free energy cannot simply be calculated by summing the local free energy at each point on the density profile. So the fluctuations around $\bar{\rho}(x)$ are correlated with fluctuations over the whole system. Here we see the kind of long-range order that has been inferred from other studies of non-equilibrium systems. Indeed, the production of entropy according to Boltzmann implies the presence of such long-range correlations. Perhaps the calculation by Derrida *et al.* is a first step to a microscopic understanding of pattern formation in non-equilibrium systems, such as the ordered

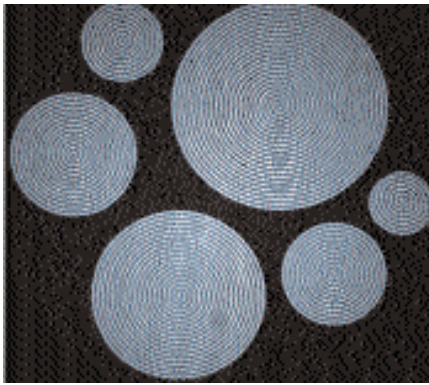


Figure 1 Pattern formation in Rayleigh–Bénard convection. Shadowgraphs of a horizontal fluid layer heated from below reveal large spiral patterns (dark corresponds to warm, rising fluid and light to cold, sinking fluid)⁷. Despite the underlying non-equilibrium nature of this process the spirals are surprisingly stable and regular.

structures seen in Rayleigh–Bénard convection of a fluid heated from below (Fig. 1).

Such striking results are emerging from

the increasing activity in studies of non-equilibrium systems^{2–6}. Our fundamental understanding of non-equilibrium states is moving forwards, but the fact that we cannot yet prove Fourier's law³ (that heat conduction is inversely proportional to the length of the conductor) from first principles shows how far we still have to go.

David Ruelle is in the Department of Mathematics, Rutgers, The State University of New Jersey, 110 Frelinghuysen Road, Piscataway, New Jersey 08854-8019, USA, and at the Institut Haute Etudes Scientifiques, 91440 Bures-sur-Yvette, France.
e-mail: ruelle@ihes.fr

1. Derrida, B., Lebowitz, J. L. & Speer, E. R. *Phys. Rev. Lett.* **87**, 150601 (2001).
2. Gallavotti, G. & Cohen, E. G. D. *Phys. Rev. Lett.* **74**, 2694–2697 (1995).
3. Eckmann, J.-P., Pillet, C.-A. & Rey-Bellet, L. *Commun. Math. Phys.* **201**, 657–697 (1999).
4. Ruelle, D. *Commun. Math. Phys.* (in the press).
5. Jaksic, V. & Pillet, C.-A. *Commun. Math. Phys.* **217**, 285–293 (2001).
6. Bertini, L., De Sole, A., Gabrielle, D., Jona-Lasinio, G. & Landim, C. *Phys. Rev. Lett.* **87**, 040601 (2001).
7. Plapp, B. B. & Bodenschatz, E. *Physica Scripta* **T67**, 111–116 (1996).

Signal transduction

Life-or-death decisions

John M. Kyriakis

In response to a protein that is linked to inflammation, cells either die or survive. Some molecular sleuthing has shed light on how the balance is tipped towards survival.

The cells of multicellular organisms are often forced to choose between living and dying. For example, immature T cells that react against the body's own proteins are made to commit cell 'suicide', a process known as programmed cell death or apoptosis. By contrast, immune cells that can fight an invading pathogen are stimulated to survive and proliferate — but only until the infection is cleared, at which point many of the now unnecessary cells also die by apoptosis. In the immune system and elsewhere, the protein tumour-necrosis factor (TNF) can signal both cell death and survival. But what determines whether a cell lives or dies in response to TNF? And is there any way in which the life and death signalling pathways can influence each other's output? Tang and colleagues¹ and De Smaele and co-workers² address these questions on pages 313 and 308 of this issue, with interesting results.

TNF is a cytokine — a protein that marshalls and coordinates immune and inflammatory responses to infections or chronic stress. TNF exerts its effects by binding to its receptors, which are present on the surfaces of many different target cells. When occupied, these receptors recruit a cohort of intracellular 'adaptor' proteins, which

couple the receptors to intracellular signalling pathways. These pathways then organize and implement the cell's response (Fig. 1, overleaf). The so-called JNK and NF- κ B signalling pathways, both of which consist of tiers of protein kinases (enzymes that add phosphate groups to target proteins), are pivotal in determining whether cells die or survive, respectively, in response to TNF^{3,4}.

The JNKs are part of the evolutionarily conserved mitogen-activated protein kinase (MAPK) family, and are implicated in cell-death pathways stimulated by environmental stresses and TNF (Fig. 1)^{3,5,6}. Once activated, JNK proteins can move from the cytoplasm of the cell into the nucleus. There, they phosphorylate and activate numerous transcription factors — proteins that control gene expression³. However, the exact mechanisms by which JNKs contribute to cell death are still unknown.

The survival pathway that is recruited by TNF requires transcription factors of the NF- κ B/Rel family (Fig. 1). In resting cells, NF- κ B is held captive outside the nucleus — away from its target genes — by proteins of the I κ B family. Binding of TNF to its receptors initiates a signalling pathway that culminates in the phosphorylation of the I κ Bs by the trimeric I κ B kinase (IKK) complex.

Phosphorylation marks I κ Bs for degradation; this liberates NF- κ B, which then heads for the nucleus to activate a programme of gene expression⁴. NF- κ B enhances cell survival by switching on genes that dampen pro-apoptotic signals; however, only a few of these genes have been identified^{4,7}.

How a cell responds to TNF can vary, and depends on the physiological context, and so it is important to understand the molecular basis on which the cell chooses between life and death. Tang *et al.*¹ and De Smaele *et al.*² were investigating in detail how the NF- κ B pathway protects cells from apoptosis. In particular, they wanted to identify additional mechanisms and target genes involved, and to determine how these genes interact with the cell-death signalling machinery. Both teams found that one of the ways in which NF- κ B protects cells from apoptosis is by substantially blunting the JNK pathway. The activation of JNK in response to TNF was much stronger and more prolonged when the NF- κ B pathway was made inoperative, as in cells lacking RelA (a subunit of NF- κ B) or IKK β (the subunit of the IKK complex whose activity is dependent on TNF), or in cells expressing a mutant I κ B that could not be degraded^{1,2}. These effects were specific to the activation of JNK by TNF: the activation of two close relatives of JNK was unimpaired, as was the activation of JNK by interleukin-1, another cytokine involved in inducing inflammation¹.

Given that NF- κ B controls gene expression, it seemed reasonable to suppose that NF- κ B might suppress the JNK pathway by activating genes that encode inhibitors of JNK activation. Tang *et al.* and De Smaele *et al.* each identified a different JNK-inhibitory gene. De Smaele *et al.*² searched for protective genes expressed specifically in cells with a functional NF- κ B pathway, not in cells in which this pathway is disrupted. They identified *gadd45 β* — one of a family of genes that encode proteins associated with cell-cycle control and DNA repair^{8,9}. They found² that the Gadd45 β protein was expressed in response to TNF in an NF- κ B-dependent way, and reduced the activation of JNK. They also found that when *gadd45 β* expression was disrupted, JNK was more strongly activated. Tang *et al.*¹ identified *xiap* as a second gene that is switched on by TNF in an NF- κ B-dependent manner. They also showed that forced expression of low levels of *xiap* reduced activation of JNK by TNF.

Just how the Gadd45 β and XIAP proteins blunt the activation of JNKs and suppress apoptosis is unclear. Curiously, each has been found before in connection with the JNKs — but as possible JNK activators. The Gadd45 proteins, expressed in response to DNA damage, can associate with and, in the case of Gadd45 γ , stimulate a protein kinase that can activate JNK⁹. So it was suggested⁹ — albeit controversially^{10,11} — that

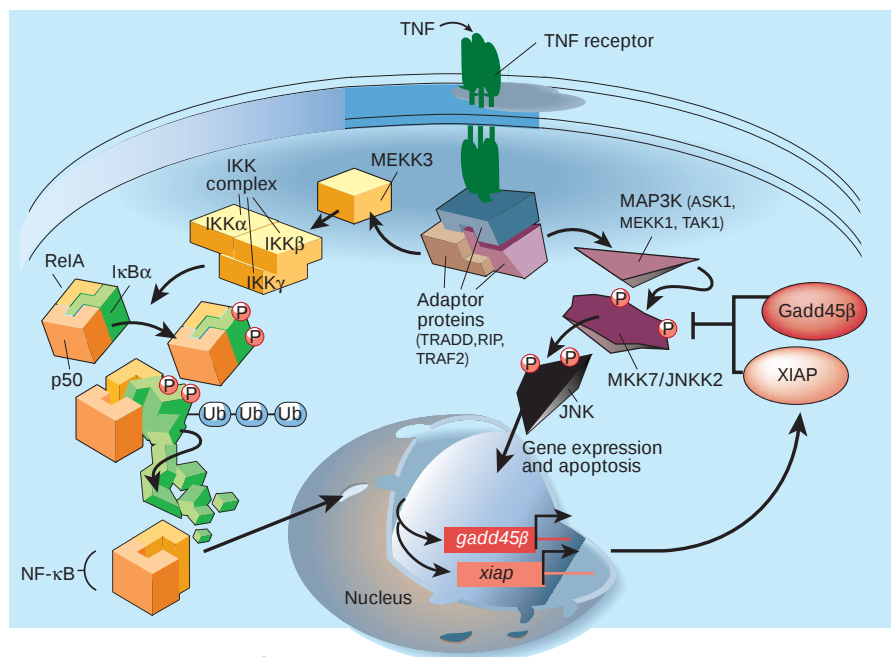


Figure 1 Crosstalk between life and death signalling pathways. The inflammatory molecule tumour-necrosis factor (TNF) recruits two pathways, which culminate in the activation of JNK or NF- κ B. Both begin (top) with the binding of TNF to its receptors, which are coupled to downstream events by adaptor proteins³. Left, in the NF- κ B pathway, the next step¹⁴ is the activation of the enzyme MEKK3. This in turn activates the IKK complex, which adds phosphate groups (P) to I κ B α , ensuring that this protein is modified with ubiquitin groups (Ub) — a prerequisite for proteolytic destruction. NF- κ B is then free to move to the nucleus, where it activates target genes. Right, in the JNK pathway, a cascade of protein kinases leads to the activation of JNKs, some of which move to the nucleus and influence the expression of a different set of genes. Tang *et al.*¹ and De Smaele *et al.*² have found that NF- κ B induces the expression of *gadd45 β* and *xiap*, which blunt the JNK pathway and so promote cell survival.

the Gadd45 proteins couple signalling pathways that are recruited by DNA damage to JNK activators. Moreover, in T lymphocytes, disruption of *gadd45 γ* dampens the activation of JNK by the T-cell antigen receptor¹². However, these results pertain primarily to Gadd45 γ , not Gadd45 β .

Meanwhile, XIAP was identified as a 'bridging' protein that indirectly couples receptors for bone morphogenetic proteins (a group of cytokines important in development) to the activation of JNK and other MAPKs³. How, then, might it blunt the activation of JNK by TNF? Perhaps, when bone morphogenetic protein is absent, XIAP instead sequesters elements that are needed for TNF to activate JNKs. One such element could be TAK1, a protein kinase that associates indirectly with XIAP and has been implicated¹³ in the activation of both IKKs and JNK by interleukin-1. Yet NF- κ B does not affect¹ the recruitment of JNK in response to interleukin-1. Further studies are clearly needed.

How might these new results^{1,2} fit into a dynamic, context-dependent model for TNF-induced signalling? Under different physiological circumstances, different adaptor proteins could be recruited to TNF receptors, favouring the activation of NF- κ B or JNK and so influencing the contribution of the 'crosstalk' effect. For example, apoptosis-

signal-regulating kinase-1 is a protein kinase required for the sustained activation of JNK by TNF in fibroblast cells⁶. Moreover, prolonged activation of NF- κ B triggers the expression of its inhibitor, I κ B, reducing further activation⁴. The contribution of these two effects might ensure that long-term exposure to TNF preferentially activates JNK-dependent apoptosis. It will be interesting to see how cell- and context-specific signalling elements are integrated with the newly discovered crosstalk between these general pathways of life and death signalling.

John M. Kyriakis is in the Diabetes Research Laboratory, Massachusetts General Hospital, 149 13th Street, Charlestown, Massachusetts 02129, USA. e-mail: kyriakis@helix.mgh.harvard.edu

1. Tang, G. *et al.* *Nature* **414**, 313–317 (2001).
2. De Smaele, E. *et al.* *Nature* **414**, 308–313 (2001).
3. Kyriakis, J. M. & Avruch, J. *Physiol. Rev.* **81**, 807–869 (2001).
4. Rothwarf, D. M. & Karin, M. http://www.stke.org/cgi/content/full/OC_sigtrans;1999/5/re1 (1999).
5. Tournier, C. *et al.* *Science* **288**, 870–874 (2000).
6. Tobiume, K. *et al.* *EMBO Rep.* **2**, 222–228 (2001).
7. Wang, C. Y. *et al.* *Science* **281**, 1680–1683 (1998).
8. Fornace, A. J. Jr, Jackman, J., Hollander, M. C., Hoffman-Liebermann, B. & Liebermann, D. A. *Ann. NY Acad. Sci.* **663**, 139–153 (1992).
9. Takekawa, M. & Saito, H. *Cell* **95**, 521–530 (1998).
10. Shaulian, E. & Karin, M. *J. Biol. Chem.* **274**, 29595–29598 (1999).
11. Wang, X., Gorospe, M., & Holbrook, N. J. *J. Biol. Chem.* **274**, 29599–29602 (1999).
12. Lu, B. *et al.* *Immunity* **14**, 583–590 (2001).
13. Wang, C. *et al.* *Nature* **412**, 346–351 (2001).
14. Yang, J. *et al.* *Nature Immunol.* **2**, 620–624 (2001).

Daedalus

Slippery light aircraft

The drag on an aeroplane, says Daedalus, is largely due to the impact of its 'wetted surface' with still air. Each colliding air molecule is momentarily adsorbed and later re-emitted at some unpredictable angle, running away with energy. The unwetted ideal, in which each molecule bounces off directly like light from a mirror, almost never happens. A rusty wing, resisting the attacking oxygen, seems worth trying; but this still leaves the nitrogen in the air. In any case, most aircraft wings are aluminium. This is superficially oxidized to transparent aluminium oxide, but is still well wetted by oxygen in the air. Surface tension is well understood in liquids, but has been largely ignored in gases. Daedalus recalls the chemical notion of anti-bonding, in which one molecule is repelled by another if the bond between them is raised by light or near-ultraviolet irradiation to an anti-bonding state.

Illuminated aircraft were actually tried during the Second World War to reduce the contrast between plane and sky as seen from the ground. Daedalus reckons that a non-wetted, anti-bonding plane should fly. DREADCO physicists are therefore fitting an aluminium plane with two sets of lights: one to raise the bond between aluminium oxide and oxygen to anti-bonding, the other to do the same for the bond between aluminium oxide and nitrogen. Such an aircraft would then remain untouched by the chief constituents of the air. Illuminated wings and fuselage should simply not be wetted.

Daedalus predicts a dramatic loss of drag. Only the windows would remain to contribute to the normal drag. The new slippery aircraft would require much less power. Even better, the propellers or the inlet fans of the jet engines could also be illuminated. The drag of these fast-moving elements would be cancelled as well, greatly increasing their efficiency. The whole plane should consume a small fraction of the usual power.

Aerodynamicists will at last be able to predict the behaviour of wings and fuselage simply. All the problems of the boundary layer will vanish: there won't be one. Molecules will bounce off an unwetted aircraft perfectly. Daedalus reckons that there should be a great saving of fuel, helping to reduce carbon dioxide emissions. Sadly, the usual painted logos must be abandoned, otherwise more light wavelengths will have to be added to reverse the bonds between the paint and the air's oxygen and nitrogen. David Jones

Benefits of female mimicry in snakes

She-male garter snakes exploit the amorous attentions of other males to warm up.

Males of several animal species mimic females either in appearance or in the chemical cues they release^{1,2}, and this mimicry has generally been interpreted in terms of alternative mating strategies — for example, a male that mimics a female may obtain stolen inseminations or avoid aggression from larger rivals³. Our studies of snakes suggest a different explanation, which relies on natural selection rather than sexual selection. Male garter snakes that produce female-like pheromones (she-males) may benefit simply because large ‘mating balls’ of amorous males form around them, transferring heat to the she-male after it emerges from hibernation and reducing its exposure to predators.

Garter snakes (*Thamnophis sirtalis parietalis*) in Manitoba, Canada, court and mate in large aggregations centred around overwintering dens. It has been suggested that the advantage of mimicry lies in a she-male's ability to confuse other males within the mating balls (which sometimes contain more than 100 males) that form around genuine females², but she-maleness has since been shown to be a transitory phase that is restricted to the first day or two after a male first emerges from his eight-month hibernation^{4,5}. The snakes are weak and slow at this time and are therefore highly vulnerable to attack by crows⁶. We have identified no mating advantage to she-males⁴, and propose an alternative explanation for female mimicry in this species.

The primary purpose of males producing female pheromones is to attract courtship from other males; indeed, most she-males are virtually obscured by the bodies of their suitors (Fig. 1). At a den near Inwood, Manitoba⁵, 49 of 53 newly emerged she-males were partly covered by other males when sighted; overall, an average of 58% of the body of each she-male was covered by other males (s.e., 4.0), whereas only 32 of 55 he-males (at more than 2 days after emergence; none of these was courted) were partly obscured, with an average of 25% cover (s.e., 4.3; one-factor ANOVA, $F_{1,106} = 31.0$, $P < 0.0001$).

She-maleness may benefit newly emerged animals (which are cold, weak and slow⁴) for two reasons. First, a snake hidden beneath other individuals may be less vulnerable to attack by crows^{6–8}. Second, courting males press vigorously against the object of their affections⁹ and may thus transfer heat — newly emerged snakes are cool (ground temperature is below 10 °C), but reproductive males are on average

warmer than 25 °C (ref. 10). Courtship could thus increase a she-male's body temperature and therefore his locomotor capacity¹¹, as well as accelerating his recovery from hibernation. Many dens are deeply shaded and smooth-sided, making it difficult for a newly emerged snake to bask in sunlight. Because reproductive males travel constantly back and forth between the den and surrounding clearings, however, a snake that attracts courtship will be covered in hot males as soon as it emerges.

Does a cold snake warm up faster if it is courted by hot males? To test this idea, we glued miniature thermal data-loggers (Thermochron I-buttons, Dallas Semiconductor) to the mid-dorsal surfaces of 24 females, which we placed in open-topped outdoor arenas measuring 1 × 1 × 1 m (ref. 4). Six arenas contained only four females each, whereas the others contained four females plus twenty males. The average temperature of females was 4 °C and that of males was 25 °C when we commenced the trials at 11:00 h to mimic newly emerging snakes. The females were heated to 20 °C within 30 min, with courted snakes heating faster than uncourted animals (repeated-measures ANOVA on data from each minute over this period, interaction $F_{29,638} = 3.87$, $P < 0.0001$).

To verify that this effect was due to heat transfer rather than to courtship-induced changes in the behaviour of courted animals, we repeated the study using dead snakes as the courtship ‘targets’, with thermistor leads implanted to measure deep-body temperature. Again, snakes exposed to courtship were heated faster than those that were not (repeated-measures ANOVA, $F_{29,203} = 1.82$, $P < 0.01$). The thermal benefit from courtship often exceeded 3 °C.

Do higher temperatures accelerate recovery from hibernation? We investigated this possibility by capturing she-males soon after they emerged, and then keeping them either warm (28 °C) or cool (10 °C). At 90-min intervals, we brought five she-males from each group to 25 °C and then held them by their tails in the den to quantify their sexual attractiveness. We scored the responses by five mate-searching males to each she-male, using a four-point scale⁴. ‘Warm’ she-males regained their he-male status within 3 hours, whereas ‘cool’ snakes remained as she-males for over 5 hours. The intensity of courtship that she-males attracted from other males therefore decreased more rapidly in snakes that were kept warmer (interaction between thermal



Figure 1 A newly emerged female-mimicking (she-male) garter snake (*Thamnophis sirtalis parietalis*), still covered in white limestone dust from its underground hibernation den, is enthusiastically courted by five other adult male snakes. The she-male may benefit through heat transfer from its suitors, as well as from reduced vulnerability to predatory crows (photo, D. O'Connor).

treatment and time, $F_{3,32} = 3.96$, $P < 0.02$).

We conclude that alternative male mating strategies such as female mimicry might have evolved through natural selection (for thermoregulation and predator defence), rather than through sexual selection, as has generally been surmised^{1–3}. She-male garter snakes may therefore manipulate their rivals' behaviour not to ‘steal’ matings, but to warm up and to reduce their own vulnerability to predation. Although intuition would favour an interpretation that female mimicry has evolved within the context of alternative mating tactics, simpler explanations should also be investigated.

R. Shine*, B. Phillips*, H. Wayet†, M. LeMaster†, R. T. Mason†

*Department of Biological Sciences, University of Sydney, New South Wales 2006, Australia
e-mail: rics@bio.usyd.edu.au

†Department of Zoology, Oregon State University, Corvallis, Oregon 97331-2914, USA

- Andersson, M. *Sexual Selection* (Princeton Univ. Press, Princeton, New Jersey, 1994).
- Mason, R. T. & Crews, D. *Nature* **316**, 59–60 (1985).
- Trivers, R. L. *Evolution* **30**, 253–269 (1976).
- Shine, R., Harlow, P. S., LeMaster, M. P., Moore, I. & Mason, R. T. *Anim. Behav.* **59**, 349–359 (2000).
- Shine, R., O'Connor, D. & Mason, R. T. *Can. J. Zool.* **78**, 1391–1396 (2000).
- Shine, R., LeMaster, M. P., Moore, I. T., Olsson, M. M. & Mason, R. T. *Evolution* **55**, 598–604 (2001).
- Olson, D. H. *Copeia* **1989**, 391–397 (1989).
- Hamilton, W. D. J. *Theor. Biol.* **31**, 295–311 (1971).
- Whittier, J. M., Mason, R. T. & Crews, D. *Behav. Ecol. Sociobiol.* **16**, 257–261 (1985).
- Shine, R., Harlow, P. S., Elphick, M. J., Olsson, M. M. & Mason, R. T. *Physiol. Biochem. Zool.* **73**, 508–516 (2000).
- Heckrotte, C. *Copeia* **1967**, 759–763 (1967).

Competing financial interests: declared none.

Antimicrobials

Peptide antibiotics in mast cells of fish

Antimicrobial peptides are increasingly recognized as a critical first line of defence against many pathogens and have been isolated from epithelial tissues and blood cells of many vertebrates, as well as from prokaryotes, plants and invertebrates^{1,2}. Here we show that ‘piscidins’, a previously undiscovered family of peptide antibiotics isolated from fish, reside in mast cells, an immune cell of uncertain function that is present in all vertebrate classes^{3,4}. Until now, no peptide antibiotic has been isolated from the mast cells of any animal, and our discovery indicates that these cells may be critical in fighting many infectious diseases.

We isolated piscidins from the tissues of an important aquacultured fish, hybrid striped bass (*Morone saxatilis* × *M. chrysops*) using a previously described method⁵ which we modified to include a further purification step after weak cation-exchange chromatography using continuous acid-urea polyacrylamide-gel electrophoresis⁶. Piscidins are 22-amino-acid peptides with a highly conserved amino terminus that is rich in histidine and phenylalanine (Table 1). The peptides probably adopt an amphipathic α-helical conformation (Fig. 1a, b).

Piscidins are more haemolytic than magainin 2, a peptide antibiotic from the aquatic frog *Xenopus laevis*, but less haemolytic than mellitin, a peptide from bee venom that has relatively weak antibacterial activity (Fig. 1c). Piscidin 3 is the least haemolytic of the piscidins; its most significant structural distinction seems to be a glycine substituted for histidine at position 17, which would tend to disrupt the amphipathic α-helix⁷ and lower

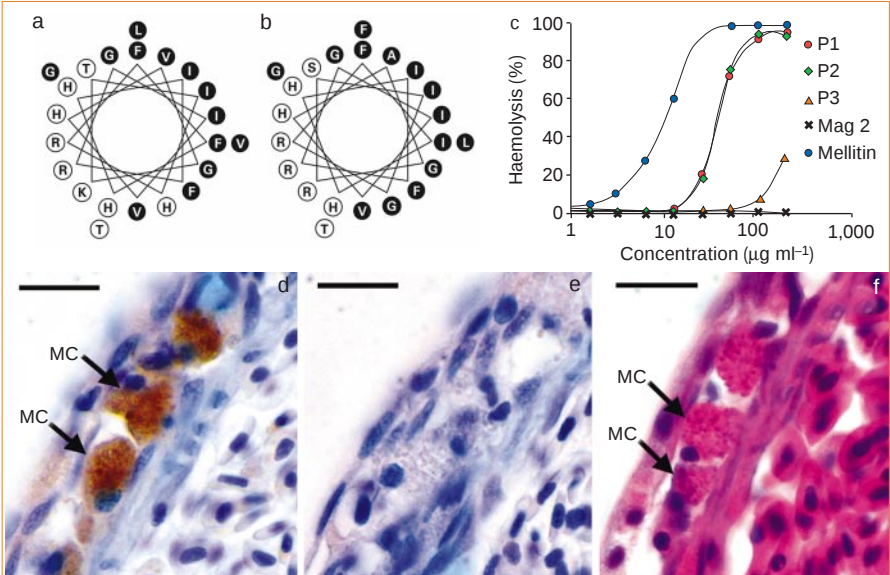


Figure 1 Characterization and localization of peptide antibiotics in hybrid striped bass. **a, b**, Helical wheels predicting amphipathic α-helical conformations for piscidins 1 (**a**) and 3 (**b**). Circles, hydrophilic residues; filled circles, hydrophobic residues. **c**, Haemolytic activity of piscidins 1–3 (P1, P2, P3), magainin 2 (Mag 2) and mellitin against human erythrocytes. **d**, Bass gill treated with anti-piscidin antibody. MC, mast cell. Affinity-purified anti-piscidin antibodies were prepared against a haemocyanin-conjugated peptide fragment (FFHHIFRGIVH)¹³. Deparaffinized tissue sections were blocked with goat serum, incubated with non-immune serum or specific antibody, and sequentially incubated with biotinylated goat anti-rabbit serum, streptavidin-conjugated peroxidase, and diaminobenzidine. **e**, Non-immune serum control of **d**. **f**, Serial section of **d**, showing mast cells with eosinophilic granules (haematoxylin and eosin stain). Scale bars: **d**, **e**, 25 μm; **f**, 10 μm.

its amphipathicity — this often correlates with reduced haemolytic activity⁸. Piscidins have potent, broad-spectrum antibacterial activity against fish pathogens, which is comparable to that of many peptides from aquatic animals⁹. This antibacterial activity seems to be correlated with haemolytic activity.

We immunolocalized piscidins to mast cells (Fig. 1d–f), the principal tissue granulocytes of vertebrates³; mast cells in gill, skin and gut, and those lining blood vessels in the viscera, were all positive. Pre-incubating anti-piscidin antibody with piscidin 1 for 30 min abrogated the signal, whereas

pre-incubation with diluent (phosphate-buffered saline) or an unrelated peptide (magainin 2) had no effect (results not shown). Not all mast cells were positive for piscidins, indicating that piscidin-negative mast cells may be at a different stage of development or may have differentiated independently of the piscidin-positive lineage. Phenotypic heterogeneity in mast cells has been reported in fish³ and mammals¹⁰.

Mast cells of members of the families Moronidae (*Morone*; Fig. 1d–f) and Sciaenidae (spot, *Leiostomus xanthurus*; croaker, *Micropogonias undulatus*; results not shown) are piscidin-positive. Both families are in the suborder Percoidei, indicating that piscidins may be evolutionarily conserved in this group, which is the largest suborder in the Perciformes, in turn the largest order of living vertebrates. The mast cells of fish species outside Percoidei are not immunohistochemically positive for piscidins.

Fish mast cells (also known as eosinophilic granule cells) are morphologically and functionally similar to their mammalian counterparts, but whether they are derived from the same lineage as mammalian mast cells is unclear. Peptide antibiotics may also be present in the mast cells of other vertebrates. Mast cells have been associated with defence against parasites for some time^{3,10}, and recent evidence suggests that they have an important role in orchestrating the activation of other effector cells, such as neutrophils, to kill pathogens¹². Our results indicate that mast cells may also

Table 1 Antibacterial activity of piscidins

Bacteria	Minimum inhibitory concentration		Minimum bactericidal concentration	
	Piscidin 1	Piscidin 3	Piscidin 1	Piscidin 3
Fish pathogens				
<i>Streptococcus iniae</i> (+)	3.1	3.1	3.1	3.1
<i>Lactococcus garviae</i> * (+)	3.1	6.3	3.1	12.5
<i>L. garviae</i> † (+)	3.1	3.1	3.1	6.3
<i>Aeromonas salmonicida</i> (–)	3.1	12.5	3.1	25.0
<i>Aeromonas hydrophila</i> (–)	0.8	1.6	0.8	1.6
<i>Vibrio alginolyticus</i> (–)	3.1	6.3	3.1	6.3
Human pathogens				
<i>Staphylococcus aureus</i> (+)	3.1	3.1	3.1	3.1
<i>Streptococcus faecalis</i> (+)	3.1	12.5	3.1	12.5
<i>Escherichia coli</i> (–)	3.1	3.1	3.1	3.1
<i>Klebsiella pneumoniae</i> (–)	3.1	6.3	3.1	6.3
<i>Shigella flexneri</i> (–)	3.1	6.3	3.1	6.3
<i>Pseudomonas aeruginosa</i> (–)	12.5	25.0	12.5	25.0

Antibacterial activity of piscidin 1 (amino-acid sequence FFHHIFRGIVHVGKTHRLVTG) and piscidin 3 (FFHHIFRGIVHAGRSIGRFLTG); conserved residues are underlined. Piscidin 2 (P2) is the same as piscidin 1, except that K is substituted for R at position 18. Mass spectrometry shows that natural piscidins sometimes have an amidated carboxy terminus. Minimum inhibitory concentration and minimum bactericidal concentration were determined using the modified Hancock method for measuring antibacterial activity of peptide antibiotics (see <http://www.cmdr.ubc.ca/bobh/MIC.htm>). Details of the methods are available at http://www.cvm.ncsu.edu/cbs/noga_ed.htm

*Resistant to oxytetracycline, kanamycin, benzylpenicillin, florfenicol, erythromycin, enrofloxacin and novobiocin.

†Resistant to oxytetracycline, erythromycin, lincomycin and doxycycline (four isolates).

participate directly in killing microbes.

Umaporn Silphaduang, Edward J. Noga

Department of Clinical Sciences, College of Veterinary Medicine, North Carolina State University, 4700 Hillsborough Street, Raleigh, North Carolina 27606, USA
e-mail: ed_noga@ncsu.edu

1. Ganz, T. & Lehrer, R. I. *Curr. Opin. Hematol.* **4**, 53–68 (1997).
2. Hancock, R. E. & Diamond, G. *Trends Microbiol.* **8**, 402–410 (2000).
3. Reite, O. B. *Fish Shellfish Immunol.* **8**, 489–513 (1998).
4. Baccari, G. C., Minucci, S., De Paulis, A. & De Santis, A. in *Mast Cells and Basophils* (eds Marone, G., Lichtenstein, L. M. & Galli, S. J.) 117–130 (Academic, New York, 2000).
5. Robinette, D. et al. *Cell. Mol. Life Sci.* **54**, 467–475 (1998).
6. Harwig, S. S., Chen, N. P., Park, A. S. & Lehrer, R. I. *Anal. Biochem.* **208**, 382–386 (1993).
7. Tossi, A., Sandri, L. & Giangaspero, A. *Biopolymers* **55**, 4–30 (2000).
8. Kondejewski, L. H. et al. *J. Biol. Chem.* **274**, 13181–13192 (1999).
9. Jia, X. et al. *Appl. Environ. Microbiol.* **66**, 1928–1932 (2000).
10. McNeill, H. P. & Austen, K. F. in *Samter's Immunologic Diseases* Vol. 1 (eds Frank, M. M., Austen, K. F., Claman, H. N. & Unanue, E. R.) 185–204 (Little, Brown and Co., Boston, 1995).
11. Wedemeyer, J., Tsai, M. & Galli, S. J. *Curr. Opin. Immunol.* **12**, 624–631 (2000).
12. Malaviya, R. & Abraham, S. N. *J. Leukoc. Biol.* **67**, 841–846 (2000).
13. Gonzalez-Cadavid, N. F. et al. *Proc. Natl Acad. Sci. USA* **95**, 14938–14943 (1998).

Pathogenesis

HLA-DQ7 antigen and resistance to variant CJD

Variant Creutzfeldt–Jakob disease (vCJD) in humans is caused by a bovine spongiform encephalopathy (BSE)-like prion strain, and so far about 100 of the many people exposed to BSE prions have developed the condition. Here we show that there is a significantly reduced frequency of the human leukocyte antigen (HLA) class-II type DQ7 in patients with vCJD, but not in those with classical CJD. This apparently protective genetic factor should aid differential diagnosis and may have important implications for understanding host susceptibility to infection by BSE prions, and the distinctive pathogenesis of vCJD, as well as in the identification of targets for prevention and therapy of vCJD.

Prion diseases are characterized by deposition of an abnormal conformation of host-encoded prion protein, PrP^{Sc}. Human prion diseases exist in inherited, sporadic and acquired forms. The idea that vCJD, which was identified in 1996, arose as a result of dietary or other exposure to the BSE agent is supported by molecular and biological strain-typing studies — it is possible that many individuals have been infected^{1–4}. The pathogenesis of vCJD is distinct from that of other forms of CJD, with a marked accumulation of PrP^{Sc} in the lymphoreticular system^{5,6}. This distinctive feature may relate to the route of infection, to host genetic factors, or to a prion-strain-specific effect⁴.

Table 1 HLA-DQ typing of CJD patients and controls

DQ type	Frequency in vCJD (50 patients)	Frequency in sporadic CJD (26 patients)	Consecutive cadaveric controls (n=197)
DQB1*02 (DQ2)	42% (21)	30.8% (8)	40.6% (80)
DQB1*04 (DQ4)	4% (2)	3.8% (1)	2.5% (5)
DQB1*05 (DQ5)	32% (16)	30.8% (8)	26.9% (53)
DQB1*06 (DQ6)	58% (29)	38.5% (10)	41.6% (82)
DQB1*0301/4/9 (DQ7)	12% (6)†	46.2% (12)	35.5% (70)
DQB1*0302/5/7/8 (DQ8)	18% (9)	15.4% (4)	19.8% (39)
DQB1*0303/6 (DQ9)	18% (9)	15.4% (4)	12.7% (25)

†P = 0.001; uncorrected from a two-sided Fisher's exact test.

The lymphoreticular system is also involved in sheep scrapie and in murine models of scrapie in which follicular dendritic cells accumulate PrP^{Sc} early in the incubation period, well before the onset of neurological signs. Prion replication in the lymphoreticular system may be a prerequisite for neuro-invasion after peripheral inoculation with low doses of infective agent⁷.

The aetiology of sporadic CJD, currently the main differential diagnosis of vCJD, remains unclear, but its random worldwide distribution and lack of clustering or association with local scrapie prevalence argues against an acquired aetiology in the majority of patients. Somatic mutation of the prion-protein gene (*PRNP*), or spontaneous formation of PrP^{Sc}, are thought to be more likely causes. A common *PRNP* polymorphism, in which either methionine or valine is encoded at codon 129, is a key determinant of genetic susceptibility to sporadic and acquired prion diseases, with heterozygosity being highly protective^{8,9}. All victims of vCJD studied so far have been methionine homozygotes at this locus⁴; about 38% of the UK population has this genotype.

In view of the distinctive pathogenesis and well-known associations between HLA and susceptibility to other infectious diseases, we compared the HLA type of vCJD patients with normal controls and with those with sporadic CJD. We carried out an initial study on 32 vCJD and 12 sporadic CJD DNA samples. Tissue samples were obtained at autopsy with the consent of relatives. These samples were HLA-typed for HLA-A, B, C, DRB and DQB1 at low-to-medium resolution using a sequence-specific-primer polymerase chain reaction (PCR-SSP) method¹⁰. All patients included were British caucasoids. HLA frequencies from a panel of 197 consecutive cadaveric British caucasoid organ donors were used as control values.

The only statistically significant deviation from control frequencies was seen at the *DQB1* locus — only 2 out of the 32 vCJD patients were positive for *DQB1**0301/4/9 (DQ7), compared with 70 out of 197 controls (35.5%), giving an uncorrected *P* value of 0.0004 (two-sided Fisher's exact test). In the case of sporadic CJD, 50% of patients were DQ7 (6 of 12), which was not significantly different from the controls (*P* = 0.36).

We then typed a further 18 vCJD and 14

sporadic CJD patients for only HLA-DQB1, using a commercial PCR-SSP kit (Reli; Dynal, UK). Table 1 shows the frequency of HLA-DQ types in 50 vCJD patients, 26 sporadic CJD patients and 197 controls. HLA-DQ types were assigned by considering the DQB1 alleles of each subject. HLA-DQ7 remained significantly under-represented in vCJD patients (Table 1). HLA-DQ frequencies in the sporadic CJD group were comparable to those of controls.

Our results suggest that the presence of DQ7 protects against vCJD, with a relative risk factor for individuals negative for DQ7 of 3.3. DQ7 typing of suspected vCJD patients, in conjunction with *PRNP* genotyping, could prove to be useful in diagnosis. Although we have typed around 50% of all reported vCJD patients, our sample size is necessarily small. As with a more prevalent disease, these data should be confirmed on a much larger group of patients, although this is not possible for vCJD at present.

A firm tissue-based diagnosis of vCJD can be achieved by tonsil biopsy⁶, but blood-based diagnostics would be preferable. Although no such test has yet been reported, combinations of blood-based markers may have strong predictive value, with DQB typing being useful for differential diagnosis and identification of high-risk groups.

The molecular basis of the protective association of DQ7 and vCJD is unknown. Class II molecules of the major histocompatibility complex (MHC) may have a direct role in disease pathogenesis, or a gene linked to the *DQB* locus may be involved. MHC class II could help in the fortuitous carriage of PrP^{Sc} around the body, or from the gut into the lymphoreticular system — with the DQ7 molecule being less efficient in this role. Alternatively, HLA-DQ7 molecules may be more effective at presenting a putative pathogenic peptide and thereby initiating an immune response. Further investigation should reveal the role of MHC class II molecules in disease pathogenesis and aid the development of secondary prophylaxis or therapeutic intervention.

Graham S. Jackson*, Jonathan A. Beck*, Cristina Navarrete†‡, Juliette Brown†, P. Margaret Sutton§, Marcela Contreras†¶, John Collinge*

*MRC Prion Unit, Institute of Neurology, University College London, Queen Square,

London WC1N 3BG, UK

e-mail: j.collinge@ic.ac.uk

†National Blood Service, Histocompatibility and

Immunogenetics, North London Centre,

Colindale Avenue, London NW9 5BG, UK

Departments of ‡Immunology, and ¶Haematology,

Royal Free and University College Medical School,

London NW3 2PF, UK

§Oxford Transplant Centre, Nuffield Department of

Surgery, Oxford Radcliffe Hospitals,

Oxford OX3 7LJ, UK

- Collinge, J., Sidle, K. C. L., Meads, J., Ironside, J. & Hill, A. F. *Nature* **383**, 685–690 (1996).
- Hill, A. F. *et al. Nature* **389**, 448–450 (1997).
- Bruce, M. E. *et al. Nature* **389**, 498–501 (1997).
- Collinge, J. *Lancet* **354**, 317–323 (1999).
- Hill, A. F. *et al. Lancet* **353**, 183–189 (1999).
- Wadsworth, J. D. F. *et al. Lancet* **358**, 171–180 (2001).
- Montresano, F. *et al. Science* **288**, 1257–1259 (2000).
- Collinge, J., Palmer, M. S. & Dryden, A. J. *Lancet* **337**, 1441–1442 (1991).
- Palmer, M. S., Dryden, A. J., Hughes, J. T. & Collinge, J. *Nature* **352**, 340–342 (1991).
- Bunce, M. *et al. Tissue Antigens* **46**, 355–367 (1995).

Brazil-nut effect

Size separation of granular particles

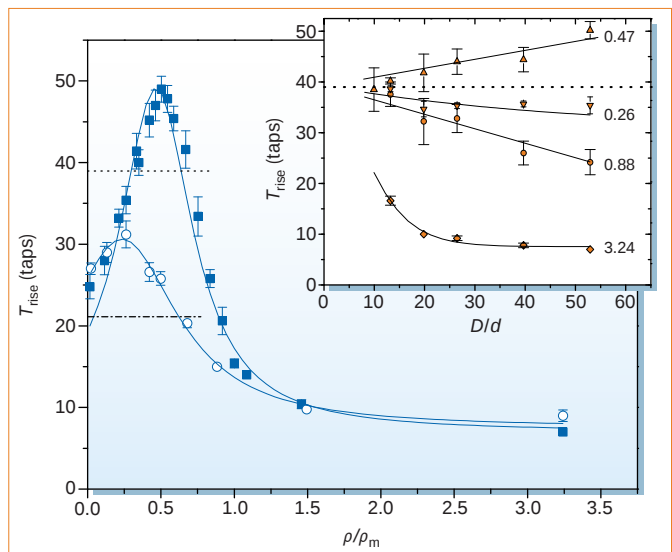
Granular media differ from other materials in their response to stirring or jostling — unlike two-fluid systems, bi-disperse granular mixtures will separate according to particle size when shaken, with large particles rising, a phenomenon termed the ‘Brazil-nut effect’^{1–8}. Mounting evidence indicates that differences in particle density affect size separation in mixtures of granular particles^{9–11}. We show here that this density dependence does not follow a steady trend but is non-monotonic and sensitive to background air pressure. Our results indicate that particle density and interstitial air must both be considered in size segregation.

Explanations of the Brazil-nut effect, which has been known since the 1930s, have focused either on infiltration of small particles into voids created underneath larger ones during shaking^{1–5} or on granular convection^{6–8}, and have implied density-independent rising times for the larger ‘intruder’ particles. However, an increase in the velocity of a large intruder with increasing density has been reported^{9,10}, suggesting that increased inertia might play a role. Furthermore, in computer simulations¹⁰, a ‘reverse’ Brazil-nut effect was found, in which groups of larger particles, if heavy enough, segregate to the bottom.

A monotonic density dependence implied by such mechanisms^{9–11} is incompatible with our measurements of intruder rising times over a wide range of size and density ratios (Fig. 1). We tracked an intruder particle in the presence of granular convection produced by vertically shaking

Figure 1 Density and size dependence of the Brazil-nut effect. The rising time, T_{rise} , of a spherical intruder of density ρ and diameter D , starting with its top at a depth $z_0 = 4.6$ cm below the surface of a vibrated granular medium consisting of $d = 0.5$ mm glass spheres ($\rho_m = 2.4$ g ml⁻¹), is plotted as a function of density ratio, ρ/ρ_m , and size ratio, D/d . Data were obtained using well-separated sinusoidal taps at normalized accelerations $\Gamma = A(2\pi f)^2/g = 5$, where A is the shaking amplitude, f is the frequency (13 Hz) and g is the Earth’s acceleration (9.81 m s⁻²). The fill height of the

container was 8.6 cm; results were similar at greater filling heights. The container diameter was 8.2 cm. A layer of glass beads, attached to the inside wall using epoxy adhesive, induces a stable, reproducible and axially symmetric convection with well-established properties^{6,7,13}. Results are shown (main panel) for fixed $D = 2.54$ cm at ambient pressure (squares) and $P = 90$ torr (circles). Dotted and dot-dashed lines show T_{rise} for tracer particles at the respective pressures in the absence of an intruder. Inset, size dependence for intruders made from four different materials (top to bottom: nylon, wood, Teflon, steel) at ambient pressure, with the densities, ρ/ρ_m , indicated to the right of the respective traces. Solid lines in the main panel are lorentzian fits, intended as visual guides.



a three-dimensional cylinder filled with smaller background particles (density, ρ_m). A spherical intruder (diameter, D ; density, ρ) was placed at a depth z_0 below the surface; a hollow acrylic ball filled with foam and lead shot was used to tune the intruder density. Material properties other than density, such as coefficients of restitution and friction, had no measurable impact.

For a fixed intruder diameter, the measured rising time, T_{rise} , to the free surface exhibits a pronounced peak as a function of ρ/ρ_m (Fig. 1). This peak is not affected by variations in shaking parameters, background medium (glass beads, poppy seeds) and system size. Compared with convection measured in the absence of an intruder (dotted line), the intruder rises faster both at large and small ρ/ρ_m , but more slowly when $\rho/\rho_m \approx 0.5$. A monotonic dependence, $T_{\text{rise}} \approx (\rho/\rho_m)^{-1/2}$, proposed for a two-dimensional system¹⁰, is incompatible with our data. The presence of a large intruder perturbs the convective flow of the background particles. Data above the horizontal dotted lines in Fig. 1 therefore do not necessarily imply sinking intruders⁹ in the absence of convection. The peak in T_{rise} becomes significant for diameter ratios $D/d > 10$, increasing with increasing intruder size (Fig. 1, inset).

Measurements of intruder velocity as a function of depth show that the increase in T_{rise} with ρ/ρ_m to the left of the peak is caused by behaviour that takes place as the particle approaches the upper surface. Deeper inside the pile, T_{rise} decreases monotonically with ρ/ρ_m . The peak is sensitive to the background air pressure, P , in the cylinder. It decreases in magnitude and shifts to lower ρ/ρ_m with

decreasing P , and vanishes as P approaches 1 torr. At this low pressure, the intruder velocity (both at the surface and within the bulk) no longer depends on ρ/ρ_m and coincides, within our resolution, with the non-zero convection velocity of the background particles in the absence of the intruder.

Our results indicate an intricate interplay between vibration-induced convection and fluidization, drag by interstitial air¹², and intruder motion. The rising time of a large intruder in a bed of smaller particles emerges as a sensitive probe of these interactions. Understanding the phenomenon described here may require a new approach that describes intruder motion in the presence of two ‘fluids’: background particles and interstitial air.

Matthias E. Möbius, Benjamin E. Lauderdale, Sidney R. Nagel, Heinrich M. Jaeger

James Franck Institute and Department of Physics, University of Chicago, Chicago, Illinois 60637, USA
e-mail: s-nagel@uchicago.edu

- Rosato, A., Strandburg, K. J., Prinz, F. & Swendsen, R. H. *Phys. Rev. Lett.* **58**, 1038–1040 (1987).
- Jullien, R. & Meakin, P. *Nature* **344**, 425–427 (1990).
- Jullien, R. & Meakin, P. *Phys. Rev. Lett.* **69**, 640–643 (1992).
- Williams, J. C. *Powder Technol.* **15**, 245–251 (1976).
- Duran, J., Rajchenbach, J. & Clement, E. *Phys. Rev. Lett.* **70**, 2431–2434 (1993).
- Knight, J. B., Jaeger, H. M. & Nagel, S. *Phys. Rev. Lett.* **70**, 3728–3731 (1993).
- Knight, J. B. *et al. Phys. Rev. E* **54**, 5726–5738 (1996).
- Cooke, W., Warr, S., Huntley, J. M. & Ball, R. C. *Phys. Rev. E* **53**, 2812–2822 (1996).
- Shinbrot, T. & Muzzio, F. J. *Phys. Rev. Lett.* **81**, 4365–4368 (1998).
- Liffman, K., Muniandy, K., Rhodes, M., Gutteridge, D. & Metcalfe, G. A. *Granular Matter* (in the press).
- Hong, D. C., Quinn, P. V. & Luding, S. *Phys. Rev. Lett.* **86**, 3423–3426 (2001).
- Pak, H. K., van Doorn, E. & Behringer, R. P. *Phys. Rev. Lett.* **74**, 4643–4646 (1995).
- Ehrichs, E. E. *et al. Science* **267**, 1632–1634 (1995).

Auxin regulates SCF^{TIR1}-dependent degradation of AUX/IAA proteins

William M. Gray^{*†}, Stefan Kepinski[‡], Dean Rouse[§], Ottoline Leyser[‡] & Mark Estelle^{*}

^{*} The Institute for Cellular and Molecular Biology, The University of Texas at Austin, Austin, Texas 78712, USA

[‡] Department of Biology, University of York, Box 373, York YO10 5YW, UK

[†] These authors contributed equally to the work

The plant hormone auxin is central in many aspects of plant development. Previous studies have implicated the ubiquitin-ligase SCF^{TIR1} and the AUX/IAA proteins in auxin response. Dominant mutations in several AUX/IAA genes confer pleiotropic auxin-related phenotypes, whereas recessive mutations affecting the function of SCF^{TIR1} decrease auxin response. Here we show that SCF^{TIR1} is required for AUX/IAA degradation. We demonstrate that SCF^{TIR1} interacts with AXR2/IAA7 and AXR3/IAA17, and that domain II of these proteins is necessary and sufficient for this interaction. Further, auxin stimulates binding of SCF^{TIR1} to the AUX/IAA proteins, and their degradation. Because domain II is conserved in nearly all AUX/IAA proteins in *Arabidopsis*, we propose that auxin promotes the degradation of this large family of transcriptional regulators, leading to diverse downstream effects.

Plant development requires the coordinated regulation of cell division, expansion and differentiation. The plant hormone indole-3-acetic acid (IAA or auxin) is fundamental in regulating many of these processes.

Genetic studies in *Arabidopsis* indicate that regulated protein degradation is required for auxin response. Recessive mutations in *AXR1* and *TIR1*, both components of the ubiquitin-mediated proteolytic pathway, result in reduced auxin response¹. *TIR1* encodes an F-box protein that interacts with the cullin AtCUL1 and a SKP1-like protein (ASK1 or ASK2) to form an SCF ubiquitin protein ligase (E3). On the basis of these results, we proposed that SCF^{TIR1} targets one or more repressors of auxin response for degradation^{2,3}. *AXR1* encodes a subunit of the enzyme that activates the ubiquitin-like protein RUB1 for conjugation to target proteins⁴. One target for RUB1 conjugation is the AtCUL1 subunit of the SCF^{TIR1} complex, and evidence suggests that modification of cullins by RUB1 is important in regulating activity of SCF ubiquitin-ligases^{5–8}.

Studies of the AUX/IAA family of transcriptional regulators have also implicated protein degradation in auxin response. The *Arabidopsis thaliana* genome contains at least 24 AUX/IAA genes, many of which were identified because of their rapid induction after auxin treatment⁹. The AUX/IAA proteins have a relative molecular mass of 20,000–35,000 (*M_r* 20K–35K) and share four conserved domains, designated I–IV. Domains III and IV mediate homo- and heterodimerization between AUX/IAA proteins and heterodimerization with members of a second large protein family called the auxin-response factors (ARFs), most of which also contain domains III and IV^{10,11}. The ARF proteins are transcription factors that bind to auxin-response elements (AuxRE) located upstream of auxin-inducible genes¹¹. Overexpression of some AUX/IAA genes was found to repress transcription of an AuxRE-reporter in transient transfection assays¹². Because the AUX/IAA proteins are not known to bind DNA, this negative regulation may occur through interaction with ARF transcription factors.

Dominant mutations conferring auxin-related phenotypes have been isolated in several AUX/IAA genes^{13–16}. These mutations all occur within the highly conserved core of domain II of each protein. Domain II has recently been demonstrated to act as a transferable protein degradation signal when fused to luciferase¹⁷. Furthermore, mutations in domain II equivalent to the dominant mutant alleles

of *AXR2/IAA7*, *AXR3/IAA17* and *SHY2/IAA3* restored stability to the luciferase fusion protein. Consistent with this finding, pulse-chase experiments reveal that the mutant *axr3-1* protein has a half-life about sevenfold greater than its wild-type counterpart¹⁸. These results indicate that rapid turnover of AUX/IAA proteins is essential for normal auxin response and that the biochemical basis for these dominant mutations is increased protein stability.

Here we show that both treatment with a proteasome inhibitor and mutations affecting the SCF^{TIR1} complex increase stability of AUX/IAA proteins. Furthermore, we demonstrate that SCF^{TIR1} physically interacts with AUX/IAA proteins. This interaction is mediated by domain II of the AUX/IAA proteins and is abolished by mutations within this motif. Auxin treatment stimulates the interaction between SCF^{TIR1} and AUX/IAA proteins and promotes their degradation. These data indicate that auxin promotes SCF^{TIR1}-dependent degradation of AUX/IAA proteins. Rapid changes in the levels of individual members of this large family of proteins are likely to result in the diverse downstream effects associated with auxin response.

Analysis of AUX/IAA stability with GUS fusions

To examine AUX/IAA protein stability, we generated transgenic plants expressing an *AXR2*–GUS (beta-glucuronidase) fusion protein under control of the cauliflower mosaic virus *CaMV* 35S promoter. Despite the presence of the *AXR2*–GUS transgene, we detected no, or in a few lines very weak, GUS activity by histochemical staining. The dominant *axr2-1* mutation results in an amino-acid substitution within the domain II motif known to be important for instability^{13,17}. When plants expressing an *axr2-1*–GUS protein were examined, abundant GUS staining was detected in the nuclei of many cells, and was especially strong in root tips (Fig. 1a). These plants exhibited several auxin-related growth phenotypes, suggesting that the GUS fusion proteins retained *AXR2* function (see below).

We employed a similar approach to compare *AXR3/IAA17* and *axr3-1* protein levels. The amino-terminal domains I and II of *AXR3* (*AXR3NT*) were fused to GUS and placed under the control of the soybean heat-shock promoter (*HS*)¹⁹. The resulting *AXR3NT*–GUS protein is non-functional but retains the bipartite nuclear localization signal spanning domains I and II. Wild-type plants expressing the *HS::AXR3NT*–GUS constructs were heat shocked at 37 °C for 2 h and stained for GUS activity 60 min after the end of the heat-shock period. Like the *AXR2*–GUS proteins, significantly more staining was detected with the *HS::axr3-1NT*–GUS construct than

[§] Present address: CSIRO Plant Industry, GPO Box 1600, Australian Capital Territory 2601, Australia.

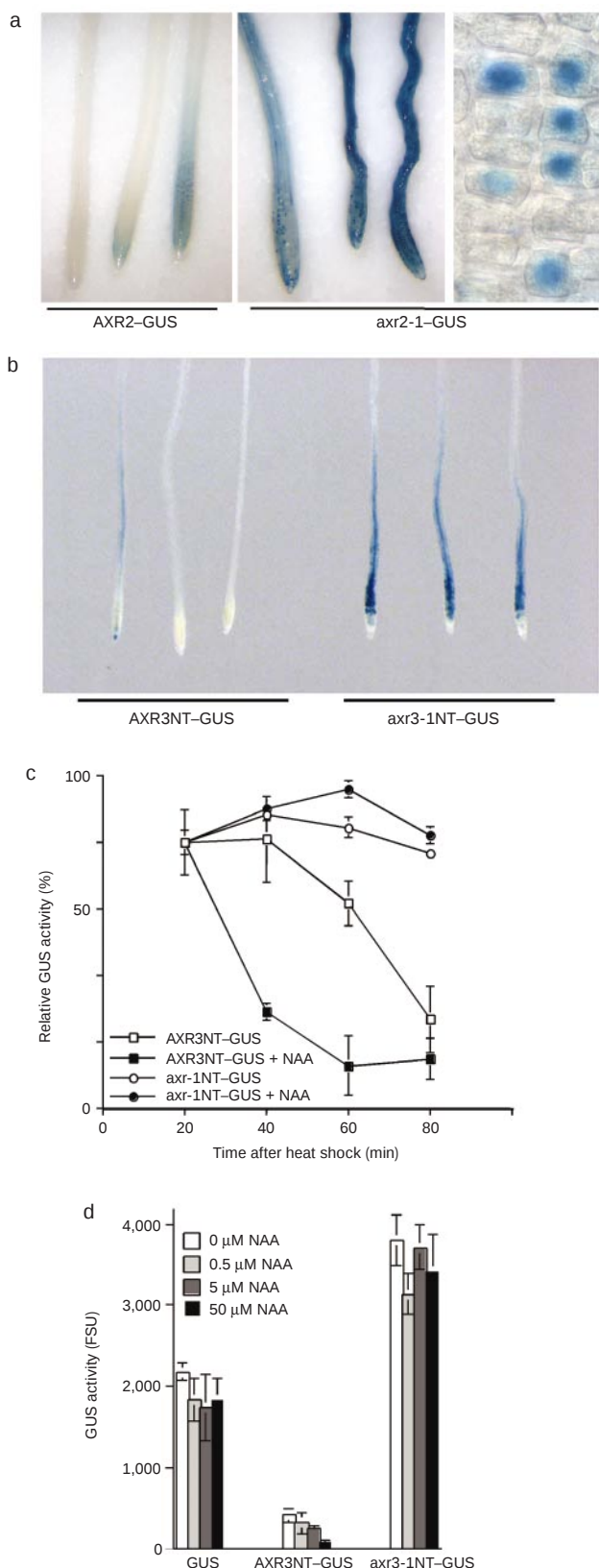


Figure 1 Analysis of AUX/IAA-GUS fusion constructs. **a**, Seven-day-old seedlings stained for GUS activity. Right, nuclear localization of *axr2-1*-GUS in root meristem cells. **b**, Seedlings were heat shocked for 2 h and stained for GUS activity 60 min after the end of the heat induction. **c**, Auxin destabilization of AXR3NT-GUS. Relative activity is expressed as percentage of the 20-min level. Error bars, s.e.m.; $n = 6$. **d**, HS::GUS, HS::AXR3NT-GUS, and HS::axr3-1NT-GUS seedlings were treated with NAA 20 min after the end of the heat-shock period. GUS activity was measured fluorometrically 50 min after NAA addition. Error bars, s.e.m.; $n = 6$.

the wild-type derivative (Fig. 1b). For both wild-type and mutant proteins, staining was primarily nuclear (data not shown). These results support previous findings suggesting that the biochemical basis for the phenotypes conferred by dominant AUX/IAA mutations is increased stability of the mutant protein^{17,18}.

We examined the possibility that auxin regulates AUX/IAA degradation using HS::AXR3NT-GUS transgenic plants. Seedlings were treated with the synthetic auxin NAA (1-naphthalene acetic acid) 20 min after the end of the heat-shock period and assayed for GUS activity at succeeding 20-min intervals. Auxin treatment promoted degradation of AXR3NT-GUS, but had no effect on *axr3-1*NT-GUS levels (Fig. 1c).

The effect of auxin on AXR3 stability was also measured in a dose-response assay. Activity of HS::AXR3NT-GUS progressively decreased as auxin concentration increased over a range of 0–50 μ M. In contrast, GUS activities of the HS::axr3-1NT-GUS and the control HS::GUS reporters were unaffected by auxin treatment over the time course of this experiment (Fig. 1d). These data indicate that auxin rapidly destabilizes the AXR3 protein and that the *axr3-1* mutation prevents this auxin-mediated degradation.

Ubiquitin-mediated degradation of AUX/IAA proteins

Because the *AXR1* and *TIR1* genes encode proteins involved in ubiquitin-mediated degradation, we tested the possibility that the ubiquitin-proteasome pathway is involved in AUX/IAA degradation. Seedlings expressing either the AXR2-GUS or AXR3NT-GUS proteins were treated with the proteasome inhibitor MG132. Both fusion proteins were stabilized by MG132 (Fig. 2). Next, we examined whether MG132 could prevent the auxin-induced degradation of AXR3NT-GUS. Heat-shocked HS::AXR3NT-GUS seedlings were treated with MG132 for 60 min, followed by a 60-min treatment with 5 μ M auxin. Histochemical staining revealed that preincubation with proteasome inhibitor largely blocked the auxin-mediated degradation of AXR3NT-GUS (Fig. 2b). In contrast,

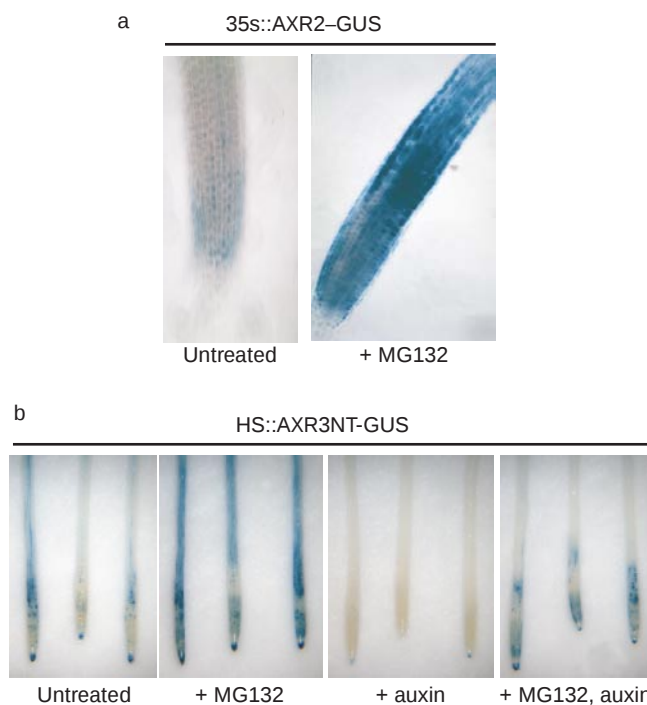


Figure 2 The proteasome inhibitor MG132 increases AUX/IAA protein stability. **a**, Seven-day-old seedlings were treated with 10 μ M MG132 for 2 h and stained for GUS activity. **b**, Nine-day-old seedlings were heat shocked for 2 h. Where indicated, seedlings were treated with 10 μ M MG132 after 1 h, and 5 μ M 2,4-D was added at the end of the heat-shock period. Sixty minutes later, seedlings were stained overnight to detect GUS activity.

MG132 treatment had no observable effect on *axr3*-INT-GUS levels over the course of the experiment (data not shown).

The effects of *axr1* and *tir1* mutations on AUX/IAA stability were investigated using the 35S::AXR2-GUS and HS::AXR3NT-GUS reporters. *tir1-1* mutants had higher AXR2-GUS levels compared with the wild type (Fig. 3a). Although the effect of the *tir1-1* mutation on AXR2-GUS levels was relatively modest, the 35S::AXR2-GUS construct conferred auxin-related defects such as leaf curling and reduced apical dominance in *tir1-1* plants. The same construct had no effect on morphology in the wild-type background. The *tir1-1*[35S::AXR2-GUS] phenotype was similar, although less severe than the phenotype of wild-type plants expressing the stabilized *axr2-1*-GUS construct, suggesting that the *tir1-1* mutation results in increased AXR2 stability (Fig. 3b).

Expression of the 35S::AXR2-GUS protein in *axr1-3* plants resulted in dramatic changes in development. Most *axr1-3*[35S::AXR2-GUS] transformants developed a single cotyledon and lacked a root meristem (Fig. 3b). Less severe transformants had fused cotyledons and a rudimentary root. These seedlings displayed greater AXR2-GUS staining compared with wild-type controls (Fig. 3a). All of the recovered *axr1-3*[35S-AXR2-GUS] transformants (>100) arrested and died before or shortly after generating the first pair of true leaves.

HS::AXR3NT-GUS levels were also elevated in *tir1* and *axr1* mutants (Fig. 3c). To examine more precisely the effects of the *axr1-12* and *tir1-1* mutations, AXR3NT-GUS levels were measured at

20-min intervals after the end of the heat-shock period. Whereas AXR3NT-GUS levels decreased rapidly in wild-type seedlings, GUS activity remained high in both *axr1-12* and *tir1-1* seedlings (Fig. 3d). AXR3NT-GUS levels were significantly higher in *axr1-12* than in *tir1-1* seedlings. This is consistent with the more severe auxin response defect exhibited by *axr1-12* plants compared with *tir1-1* plants (Fig. 3c and data not shown).

To confirm that the GUS fusion proteins accurately reflected protein stability, polyclonal antiserum was raised against AXR2 and used to examine protein levels. Although the antiserum detected recombinant AXR2 from *Escherichia coli* extracts, we were unable to detect AXR2 clearly in plant extracts on protein blots. As an alternative approach, [³⁵S]-methionine/cysteine was used to metabolically label seedling proteins. The AXR2 antiserum immunoprecipitated a 29K protein that co-migrated with the recombinant AXR2 protein (Fig. 3e). When immunoprecipitations were performed using extracts prepared from plants expressing an *axr2-1*-GFP (green fluorescent protein) fusion protein, an additional 59K species immunoprecipitated, suggesting that the antiserum does indeed recognize the AXR2 protein (Fig. 3e). Substantially more AXR2 protein was detected in *axr2-1* plants than in wild-type plants (Fig. 3e). Also consistent with the reporter analysis, more AXR2 protein was immunoprecipitated from *axr1-12* seedlings than in the wild type. In pulse-chase experiments (Fig. 3f), AXR2 exhibited a half-life of 10.8 ± 1.1 min in wild type compared with 28 ± 3.9 min in *axr1-12*. These findings validate the results obtained with the

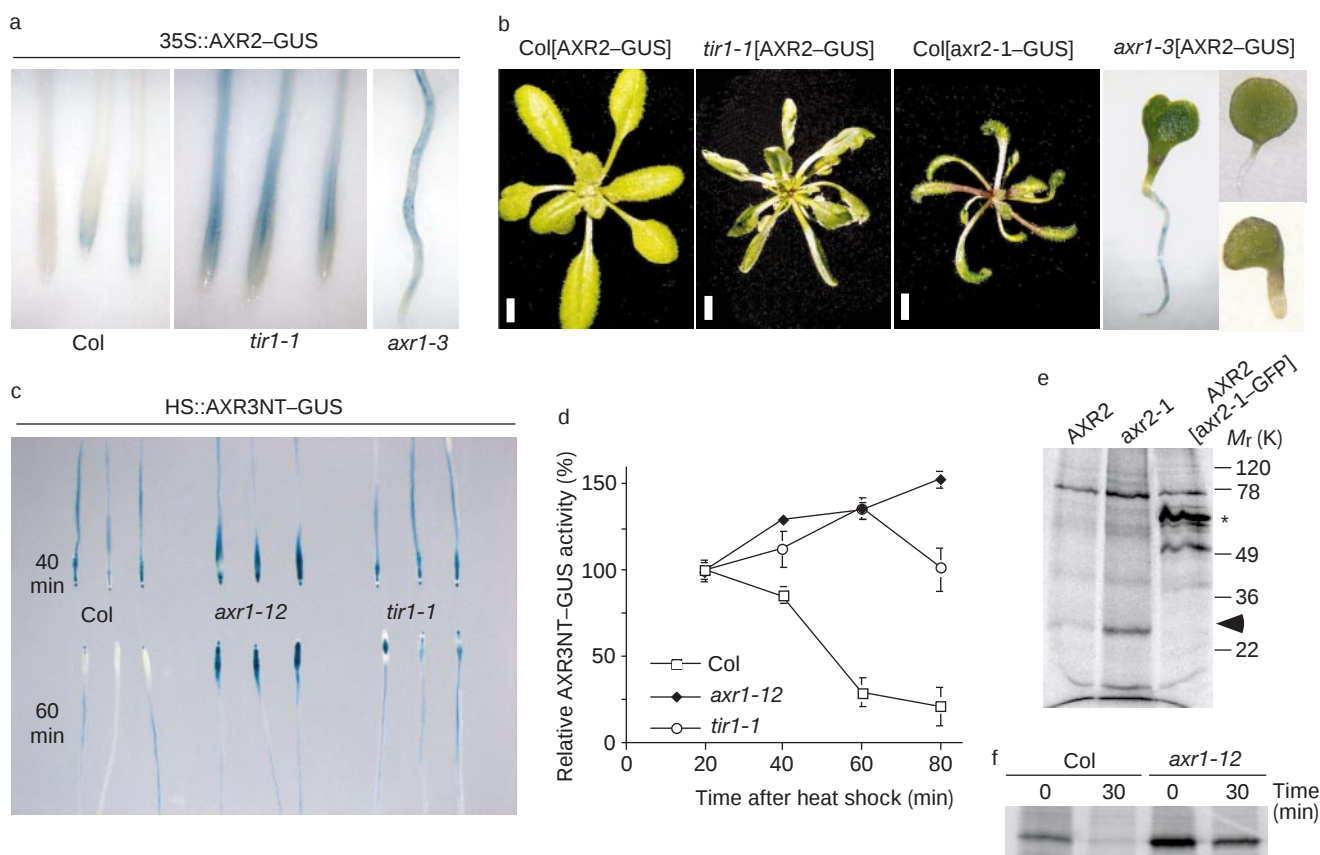


Figure 3 AUX/IAA proteins exhibit increased stability in *axr1* and *tir1* mutants. **a**, GUS histochemical staining of seedlings containing the 35S::AXR2-GUS reporter. **b**, Twenty-five-day-old plants. *tir1-1* plants are indistinguishable from the Col[AXR2-GUS] plant shown. Scale bars, 5 mm. Right, 6-day-old primary transformants. The seedling on the left was stained for GUS activity. **c**, Seedlings were stained for GUS activity 40 (top) or 60 (bottom) min after the end of the heat-shock induction. **d**, AXR3NT-GUS activity at 20-min intervals after heat shock. Relative activity is expressed as percentage of the

20-min level. Error bars, s.e.m.; $n = 6$. **e**, Proteins were extracted from [³⁵S]-methionine/cysteine-labelled 7-day-old seedlings and AXR2 protein immunoprecipitated with the anti-AXR2 antibody. The arrowhead and asterisk indicate the positions of the AXR2 and *axr2-1*-GFP proteins, respectively. **f**, AXR2 was immunoprecipitated from metabolically labelled seedlings immediately after the labelling period (0 min) or 30 min after chasing with 1 mM methionine/cysteine.

GUS reporters and suggest that AXR2 and AXR3 are targeted for ubiquitin-mediated proteolysis by the SCF^{TIR1} ubiquitin-ligase.

AUX/IAA proteins interact with SCF^{TIR1}

Our results suggest that AUX/IAA protein turnover is dependent on SCF^{TIR1}. To examine whether SCF^{TIR1} physically interacts with AUX/IAA proteins, the AXR2 antibody was used in immunoprecipitation experiments with extracts prepared from seedlings expressing the *c-myc* epitope-tagged TIR1 derivative. TIR1-Myc was readily detected in anti-AXR2 immunoprecipitates but was absent from control precipitations using the AXR2 pre-immune serum (Fig. 4a).

We explored the interaction between SCF^{TIR1} and the AUX/IAA proteins further using *in vitro* pull-down assays. Recombinant glutathione S-transferase (GST)-AXR2 was synthesized in *Escherichia coli* and purified with the GST tag. Purified protein was incubated with crude lysate prepared from *Arabidopsis* seedlings, repurified, and immunoblotted with *c-myc*, AtCUL1 and ASK2 antibodies. TIR1-Myc and AtCUL1 both co-purified with the GST-AXR2 fusion protein but were absent in control pull-down assays using GST alone (Fig. 4b, outer lanes). The Skp1-like proteins ASK1 and ASK2 were also present in GST-AXR2 pull-down assays. Because the ASK2 antibody cross-reacts with the

co-migrating GST protein, we could not confirm that ASK1 and ASK2 were missing from the GST control.

We examined the effect of the *axr2-1* mutation on interaction with SCF^{TIR1} using a GST-*axr2-1* mutant derivative. SCF^{TIR1} did not co-purify with the mutant protein, indicating that the single-base-pair *axr2-1* mutation prevents the protein from interacting with the SCF complex (Fig. 4b, centre lane).

To determine whether SCF^{TIR1} interacts with additional AUX/IAA proteins, we tested GST-AXR3 in pull-down assays. Similar to the results obtained with AXR2, GST-AXR3 co-purified with TIR1 protein and the *axr3-1* mutation substantially disrupted this interaction (Fig. 4c).

Because the *axr2-1* and *axr3-1* mutations disrupt interaction with SCF^{TIR1}, we tested whether domain II functions as a TIR1 interaction domain. A truncated derivative of the GST-AXR2 fusion protein containing only domains I and II was capable of interacting with TIR1 in a pull-down assay. Similarly, TIR1 was able to interact, albeit at a reduced level, with a GST-AXR2 fusion protein containing only domain II (AXR2⁷¹⁻¹⁰⁰). In contrast, when a short deletion was introduced into the highly conserved core of domain II, this mutant derivative of AXR2 (AXR2^{Δ86-88}) interacted very weakly with TIR1 (Fig. 4d). These data demonstrate that domain II is both necessary and sufficient to bind SCF^{TIR1}.

Because auxin promoted degradation of the AXR3-GUS fusion protein (Fig. 1), we examined the possibility that auxin regulates the interaction between AUX/IAA proteins and SCF^{TIR1}. We performed GST pull-down assays with AXR2 and AXR3 fusion proteins using crude *Arabidopsis* extracts prepared from seedlings treated with the synthetic auxin 2,4-D (2,4-dichlorophenoxy acetic acid) before protein extraction. Pull-down assays with extracts prepared from auxin-treated plants yielded more TIR1-Myc protein than control assays using untreated extracts. This increase was apparent after treatments as short as 5 min, increased until at least 60 min, and declined by 240 min (Fig. 5a; top, middle). Western blot analysis confirmed that this increase was not due to an increase in TIR1-Myc abundance in the extracts prepared from auxin-treated plants (Fig. 5a; bottom).

Applied auxin also promoted the SCF^{TIR1}-AUX/IAA interaction in a dose-dependent manner (Fig. 5b). Auxin treatment enhanced the SCF^{TIR1}-AXR2/AXR3 interaction at concentrations as low as 0.5 μM. This dose-response relationship correlates well with the effects of increasing concentrations of auxin on AXR3NT-GUS stability.

Discussion

Previous genetic and biochemical studies implicated SCF^{TIR1} as a positive regulator of auxin response in *Arabidopsis*². We have proposed that SCF^{TIR1} promotes auxin response by targeting one or more negative regulators of the pathway for ubiquitin-mediated

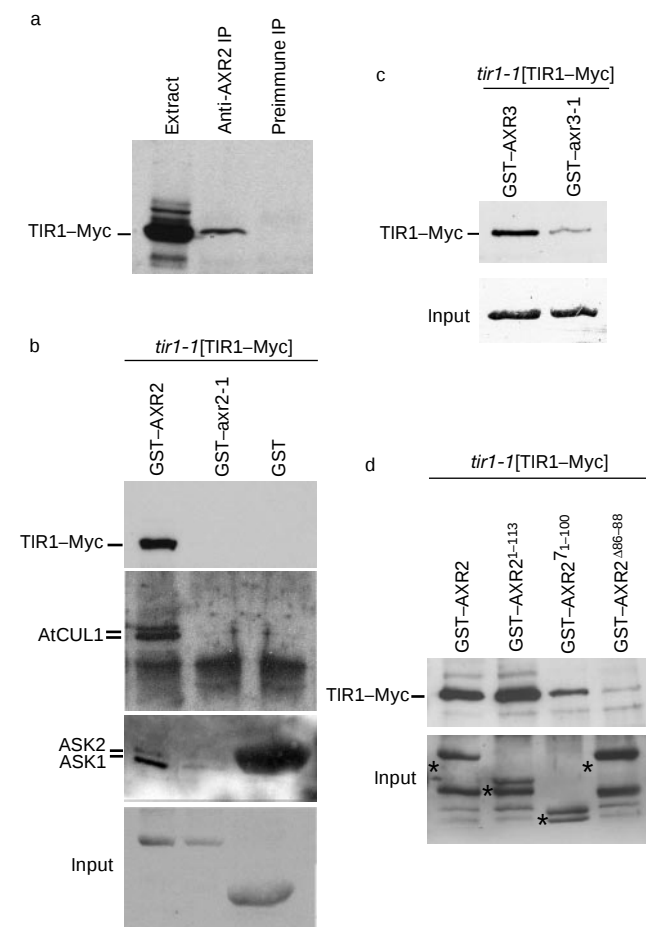


Figure 4 SCF^{TIR1} interacts with AUX/IAA proteins. **a**, Immunoprecipitates (IP) were blotted and probed with the anti-*c-myc* antibody. **b**, Recombinant GST-AXR2, GST-*axr2-1*, and GST proteins were used in pull-down assays with extracts prepared from *tir1-1*[TIR1-Myc] seedlings. Pull-down assays were immunoblotted with the indicated antibodies. The anti-ASK2 antibody detects both the ASK1 and ASK2 proteins. **c**, GST-AXR3 and GST-*axr3-1* pull-down assays were probed with anti-*c-myc* antibody. **d**, Pull-down assays with full-length GST-AXR2 and deletion derivatives were immunoblotted and probed with anti-*c-myc* antibody. Input GST-AXR2 protein is visualized in the lower panel. The position of full-length protein is indicated with an asterisk in each lane.

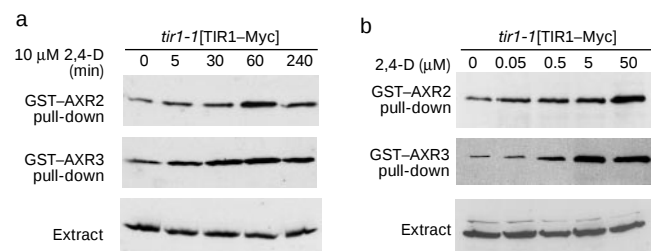


Figure 5 Auxin promotes the SCF^{TIR1}-AUX/IAA interaction. **a**, Pull-down assays were performed with extracts prepared from 6-day-old seedlings treated with 10 μM 2,4-D as indicated, and TIR1-Myc was detected by immunoblotting. Anti-*c-myc* western blots of the extracts confirmed that auxin treatment does not affect TIR1-Myc abundance (bottom). **b**, As in **a** except seedlings were treated with increasing concentrations of 2,4-D for 60 min.

degradation¹. In this report we demonstrate that the AUX/IAA proteins are targeted for degradation by SCF^{TIR1} in response to auxin. Stabilization of the AUX/IAA proteins, either by recessive mutations that affect the SCF or by dominant mutations in the AUX/IAA genes, causes dramatic defects in auxin response and morphology. These results provide a mechanistic link between the genetically defined AXR1–TIR1 pathway and two families of transcriptional regulators, the AUX/IAA and ARF proteins.

We suggest that AUX/IAA proteins are substrates of SCF^{TIR1} and the domain II of AUX/IAA proteins functions as a signal that targets these proteins for degradation^{17,18}. Our reporter and immunological findings support this hypothesis because the *axr2-1* and *axr3-1* mutations resulted in increased protein stability. In addition, mutations in *TIR1* or *AXR1*, or treatment with the proteasome inhibitor MG132, caused increased AXR2 and AXR3 stability suggesting that SCF^{TIR1} ubiquitinates these proteins, marking them for degradation by the 26S proteasome.

Our data suggest that domain II destabilizes AXR2 and AXR3 by targeting them to SCF^{TIR1}. Auxin causes reduced protein stability by promoting this interaction whereas the dominant AUX/IAA mutations confer increased protein stability by preventing the interaction between AUX/IAA proteins and the SCF. Although we have demonstrated SCF^{TIR1}–AUX/IAA interaction in crude extracts, we were unable to detect an interaction between recombinant AXR2 or AXR3 and immunopurified SCF^{TIR1}. This indicates that the plant extract provides a factor that facilitates the interaction. It is possible that SCF^{TIR1}–AUX/IAA binding is regulated by phosphorylation, as several yeast and mammalian SCF substrates must be phosphorylated to interact with their cognate SCFs²⁰. Indeed, a MAP kinase activity was recently identified that is rapidly and transiently induced by auxin²¹. However, domain II, shown here to be necessary and sufficient for SCF^{TIR1} recognition, does not contain any conserved sites of phosphorylation²². It is possible that an additional protein, serving as a bridge between the SCF and substrate, is phosphorylated in response to auxin. Alternatively, domain II may be subject to a different type of post-translational modification. The identification of the modification and/or cofactor that is required for SCF^{TIR1} binding will provide important insight into the upstream events in the auxin-response pathway.

Our understanding of AUX/IAA protein function is based largely on genetic studies. The phenotypes of the gain-of-function *axr2*, *axr3*, *shy2* and *iaa28* mutations illustrate the consequences of failure to degrade individual members of the family. In general, accumulation of each protein results in decreased auxin response. For example, the *axr2-1* mutant is deficient in auxin induction of all members of the AUX/IAA gene family, indicating that stabilized AXR2 represses transcription of these genes⁹. In addition, transfection experiments demonstrate that some AUX/IAA proteins repress auxin-dependent gene expression¹². However, it is important to note that some aspects of the *axr3-1* phenotype are more consistent with auxin hypersensitivity, suggesting that individual members of the family may have positive effects on auxin response.

Although the mechanism by which AUX/IAA proteins affect auxin response is unknown, one simple possibility is that they prevent the formation of ARF protein dimers. The ARF proteins seem to bind palindromic auxin response elements and activate transcription more efficiently as dimers^{23,24}. Because AUX/IAA proteins can heterodimerize with ARFs, they may act by preventing formation of active ARF dimers¹¹. In the case of ARFs that activate transcription, this will result in repression of transcription. For those ARFs that seem to act as repressors (for example, ARF1), interaction with an AUX/IAA protein could have a positive effect on transcription of target genes. This view is strongly supported by the effects of overexpression of the AXR2–GUS fusion protein in *axr1-3* plants. *axr1-3*[AXR2–GUS] transformants developed with single or fused cotyledons strikingly similar to loss-of-function mutants of the ARF transcription factor MP (ref. 25). Thus it is likely that

increased AXR2–GUS levels in *axr1-3* plants repress the ability of MP to regulate auxin-responsive genes.

The *tir1* and *axr1* mutations probably have a global effect on AUX/IAA stability. Domain II is conserved in 24 of the 29 members of the family, so it is likely that most of these proteins are degraded in an auxin-dependent manner. In this context, it is important to note that the *Arabidopsis* genome encodes several proteins with high homology to TIR1 as well as one AXR1-like protein. These related gene products are likely to be at least partially redundant with TIR1 and AXR1. This would explain the relatively modest effect of the *tir1-1* mutation on AUX/IAA stability and auxin response in general.

On the basis of the results presented in this study, we propose the following model for auxin response (Fig. 6). Basal levels of AUX/IAA proteins repress the auxin-response pathway. Auxin derepresses the pathway by promoting AUX/IAA binding to SCF^{TIR1} and related SCF complexes, leading to their degradation. SCF^{TIR1} function requires AXR1-dependent RUB1 modification of the AtCUL1 subunit of the SCF. AUX/IAA proteolysis results in a transient derepression of the pathway until new AUX/IAA proteins can be synthesized and restore repression. According to this model, auxin-induced expression of the AUX/IAA genes is a negative-feedback loop that ensures tight regulation of the response similar to the rapid NF- κ B activation of its inhibitor, I κ B²⁶. Auxin-induced destabilization of the AUX/IAA proteins would permit the formation of ARF–ARF dimers and hence a higher level of transcription of auxin-regulated genes. Auxin is known to elicit a diverse array of responses during the plant's life cycle. The key to this complexity may lie in the differences in expression of AUX/IAA family members, as well as differences in degradation kinetics. Indeed, the limited data available suggest striking differences in the instability of AUX/IAA proteins^{18,27}. Consistent with this possibility, we find that AXR2 interacts with SCF^{TIR1} much more efficiently than AXR3 (W.M.G. and M.E., unpublished data), which may account for the shorter half-life of the AXR2 protein (Fig. 3f)¹⁸. Extrapolated to the

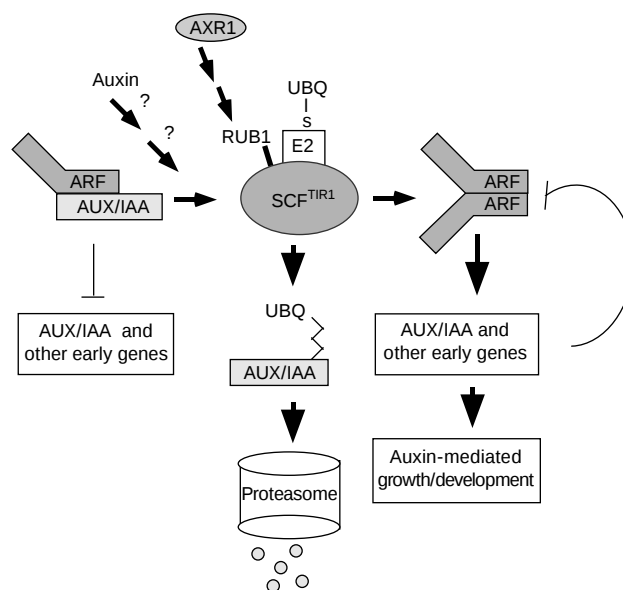


Figure 6 Model for auxin response. AUX/IAA proteins repress the auxin-response pathway by negatively regulating ARF transcription factors. Auxin promotes the ubiquitination of AUX/IAA proteins by targeting them to the SCF^{TIR1} ubiquitin-ligase. The subsequent degradation of AUX/IAA proteins results in activation of ARF and derepression of the auxin-response pathway. Because AUX/IAA genes themselves are rapidly induced by auxin, a negative-feedback loop exists with the newly synthesized AUX/IAA proteins restoring repression upon the pathway. Although the mechanism of AUX/IAA action is unclear, one possible mechanism is by preventing the formation of ARF–ARF dimers.

entire family, this would lead to considerable temporal variation in the relative abundance of individual AUX/IAA proteins in response to an auxin pulse. Such dynamics may account for the diversity of auxin responses observed in the plant. □

Methods

Plant material

All lines employed in this study were in the Columbia ecotype. Seedlings were grown under sterile conditions on vertically oriented ATS plates²⁸. Seedlings used for protein extractions were grown for 5–7 d in liquid ATS media.

Reporter constructs

We fused the 400-base pair (bp) fragment of the soybean heat-shock promoter HS6871 (ref. 19) N-terminally to *GUS* (*HS::GUS*), domains I and II of *AXR3* and *GUS* (*HS::AXR3NT-GUS*), and domains I and II of *axr3-1* and *GUS* (*HS::axr3-1NT-GUS*) using the vector pB101.3 (Clontech). The *AXR2* coding sequence was cloned into the *Bam*HI site of pB1121 (Clontech).

Heat induction and GUS assays

Seedlings were submerged in liquid ATS and heat shocked for 2 h at 37 °C. Plants were sampled at 20, 40, 60 and 80 min thereafter and stored in liquid nitrogen until protein extraction, or in the case of histochemical reactions, assayed immediately. Auxin treatments were performed by adding NAA 20 min after the end of the heat-shock period. GUS activity was measured as previously described²⁹. Fluorometric assays were performed by incubating sample extracts in 2 mM MUG (4-methylumbelliferyl- β -D-glucuronide), 50 mM KPO₄ (pH 7.0), 0.1% Sarkosyl (BDH), 0.1% Triton X-100, 10 mM β -mercaptoethanol and 10 mM EDTA for 16 h followed by analysis with a Dynex MFX microtitre plate fluorometer. Extracts were prepared from ten seedlings and data were normalized against total protein levels.

Antibodies

The *AXR2* coding sequence was cloned in-frame into the *Bam*HI site of the GST fusion vector pGEX-2T and introduced into *E. coli* strain MC1061. Stationary phase cells were diluted tenfold and grown for 1 h at 30 °C before induction with 0.1 mM IPTG. Cells were collected after 4 h of growth, resuspended in PBS buffer with 0.5% Triton X-100, and lysed by sonication. The GST-*AXR2* fusion protein was purified and subjected to SDS-PAGE, excised from the gel and injected into a rabbit to generate anti-*AXR2* antisera (Cocalico Biologicals). Crude antiserum was affinity purified against nitrocellulose-bound GST-*AXR2* fusion protein³⁰. Anti-*c-myc* monoclonal antibody was purchased from BabCo. The anti-ASK2 and anti-AtCUL1 polyclonal antibodies have been previously described².

Immunoprecipitations and pull-down assays

Immunoprecipitations were performed as previously described². For GST-*AXR2* and GST-*AXR3* pull-down assays, 4 μ g of purified fusion protein was added to 2.5 mg of crude *Arabidopsis* extract prepared from 7-day-old seedlings. Extracts were prepared by homogenizing seedlings in Buffer C (ref. 2) supplemented with 1 mM dithiothreitol, 10 μ M MG132, 10 mM β -glycerolphosphate, 1 mM NaF and 1 mM orthovanadate. The resulting homogenate was cleared by microcentrifugation for 15 min. Where indicated, seedlings were treated with 2,4-D before extraction. Following addition of the glutathione-agarose-bound GST fusion protein, extracts were incubated at 4 °C with gentle agitation for 3 h. Glutathione beads were collected by brief centrifugation, washed three times in the above buffer, resuspended in SDS-PAGE sample buffer and subjected to SDS-PAGE electrophoresis and immunoblotting.

Metabolic labelling

Seven-day-old seedlings were transferred to 4 ml of ATS medium containing 200 mCi ³⁵S-Trans label (ICN) and grown for 3.5 h. Labelled seedlings were washed and proteins extracted immediately or after a 30-min chase in medium containing 1 mM methionine/cysteine and 100 μ g ml⁻¹ cycloheximide. *AXR2* was immunoprecipitated with affinity-purified anti-*AXR2* antibody as described above. *AXR2* half-life ($t_{1/2}$) was calculated using the formula $t_{1/2} = 0.693t/\ln(N_0/N_x)$, where t is time in minutes. N_0 and N_x equal the amounts of *AXR2* at $t = 0$ and $t = 30$ min, respectively. Values presented are the mean of three independent experiments ($\pm$ s.d.)

Received 17 July; accepted 13 September 2001.

- Gray, W. M. & Estelle, I. Function of the ubiquitin-proteasome pathway in auxin response. *Trends Biochem. Sci.* **25**, 133–138 (2000).
- Gray, W. M. *et al.* Identification of an SCF ubiquitin-ligase complex required for auxin response in *Arabidopsis thaliana*. *Genes Dev.* **13**, 1678–1691 (1999).
- Ruegger, M. *et al.* The TIR1 protein of *Arabidopsis* functions in auxin response and is related to human SKP2 and yeast grr1p. *Genes Dev.* **12**, 198–207 (1998).
- del Pozo, J. C., Timpte, C., Tan, S., Callis, J. & Estelle, M. The ubiquitin-related protein RUB1 and auxin response in *Arabidopsis*. *Science* **280**, 1760–1763 (1998).
- Lyapina, S. *et al.* Promotion of NEDD8-CUL1 conjugate cleavage by COP9 signalosome. *Science* **292**, 1382–1385 (2001).
- Osaka, F. *et al.* Covalent modifier NEDD8 is essential for SCF ubiquitin-ligase in fission yeast. *EMBO J.* **19**, 3475–3484 (2000).
- Podust, V. N. *et al.* A Nedd8 conjugation pathway is essential for proteolytic targeting of p27^{Kip1} by ubiquitination. *Proc. Natl Acad. Sci. USA* **97**, 4579–4584 (2000).
- del Pozo, J. C. & Estelle, M. The *Arabidopsis* cullin AtCUL1 is modified by the ubiquitin-related protein RUB1. *Proc. Natl Acad. Sci. USA* **96**, 15342–15347 (1999).
- Abel, S., Nguyen, M. D. & Theologis, A. The PS-IAA4/5-like family of early auxin-inducible mRNAs in *Arabidopsis thaliana*. *J. Mol. Biol.* **251**, 533–549 (1995).
- Kim, J., Harter, K. & Theologis, A. Protein-protein interactions among the Aux/IAA proteins. *Proc. Natl Acad. Sci. USA* **94**, 11786–11791 (1997).
- Ulmasov, T., Hagen, G. & Guilfoyle, T. J. ARF1, a transcription factor that binds to auxin response elements. *Science* **276**, 1865–1868 (1997).
- Ulmasov, T., Murfett, J., Hagen, G. & Guilfoyle, T. J. Aux/IAA proteins repress expression of reporter genes containing natural and highly active synthetic auxin response elements. *Plant Cell* **9**, 1963–1971 (1997).
- Nagpal, P. *et al.* *AXR2* encodes a member of the Aux/IAA protein family. *Plant Physiol.* **123**, 563–574 (2000).
- Rogg, L. E., Lasswell, J. & Bartel, B. A gain-of-function mutation in *iaa28* suppresses lateral root development. *Plant Cell* **13**, 465–480 (2001).
- Rouse, D., Mackay, P., Stirnberg, P., Estelle, M. & Leyser, O. Changes in auxin response from mutations in an AUX/IAA gene. *Science* **279**, 1371–1373 (1998).
- Tian, Q. & Reed, J. W. Control of auxin-regulated root development by the *Arabidopsis thaliana* SHY2/IAA3 gene. *Development* **126**, 711–721 (1999).
- Worley, C. K. *et al.* Degradation of Aux/IAA proteins is essential for normal auxin signalling. *Plant J.* **21**, 553–562 (2000).
- Ouellet, F., Overvoorde, P. J. & Theologis, A. IAA17/AXR3. Biochemical insight into an auxin mutant phenotype. *Plant Cell* **13**, 829–842 (2001).
- Smart, C. M., Scofield, S. R., Bevan, M. W. & Dyer, T. A. Delayed leaf senescence in tobacco plants transformed with TMR, a gene for cytokinin production in *Agrobacterium*. *Plant Cell* **3**, 647–656 (1991).
- Deshais, R. J. SCF and Cullin/Ring H2-based ubiquitin ligases. *Annu. Rev. Cell Dev. Biol.* **15**, 435–467 (1999).
- Mockaitis, K. & Howell, S. H. Auxin induces mitogenic activated protein kinase (MAPK) activation in roots of *Arabidopsis* seedlings. *Plant J.* **24**, 785–796 (2000).
- Ramos, J. A., Zenser, N., Leyser, H. M. & Callis, J. Rapid degradation of Aux/IAA proteins requires conserved amino acids of domain II and is proteasome-dependent. *Plant Cell* **13**, 2349–2360 (2001).
- Ulmasov, T., Hagen, G. & Guilfoyle, T. J. Dimerization and DNA binding of auxin response factors. *Plant J.* **19**, 309–319 (1999).
- Ulmasov, T., Hagen, G. & Guilfoyle, T. J. Activation and repression of transcription by auxin-response factors. *Proc. Natl Acad. Sci. USA* **96**, 5844–5849 (1999).
- Berleth, T. & Jurgens, G. The role of the *monopteros* gene in organising the basal body region of the *Arabidopsis* embryo. *Development* **118**, 575–587 (1993).
- Karin, M. & Ben-Neriah, Y. Phosphorylation meets ubiquitination: the control of NF- κ B activity. *Annu. Rev. Immunol.* **18**, 621–663 (2000).
- Abel, S., Oeller, P. W. & Theologis, A. Early auxin-induced genes encode short-lived nuclear proteins. *Proc. Natl Acad. Sci. USA* **91**, 326–330 (1994).
- Lincoln, C., Britton, J. H. & Estelle, M. Growth and development of the *axr1* mutants of *Arabidopsis*. *Plant Cell* **2**, 1071–1080 (1990).
- Stomp, A.-M. in *GUS Protocols* (ed. Gallagher, S. R.) 103–113 (Academic, London, 1991).
- Pringle, J. R. *et al.* Fluorescence microscopy methods for yeast. *Methods Cell Biol.* **31**, 357–435 (1989).

Acknowledgements

This work was supported by grants from the National Institutes of Health (M.E.), the US Department of Energy (M.E.), the Texas Higher Education Coordinating Body Advance Research Program (M.E.) and the UK Biotechnology and Biological Science Research Council (O.L.).

Correspondence and requests for materials should be addressed to M.E. (e-mail: mestelle@mail.utexas.edu).

A histone H3 methyltransferase controls DNA methylation in *Neurospora crassa*

Hisashi Tamaru & Eric U. Selker

Institute of Molecular Biology, University of Oregon, Eugene, Oregon 97403-1229, USA

DNA methylation is involved in epigenetic processes such as X-chromosome inactivation, imprinting and silencing of transposons. We have demonstrated previously that *dim-2* encodes a DNA methyltransferase that is responsible for all known cytosine methylation in *Neurospora crassa*. Here we report that another *Neurospora* gene, *dim-5*, is required for DNA methylation, as well as for normal growth and full fertility. We mapped *dim-5* and identified it by transformation with a candidate gene. The mutant has a nonsense mutation in a SET domain of a gene related to histone methyltransferases that are involved in heterochromatin formation in other organisms. Transformation of a wild-type strain with a segment of *dim-5* reactivated a silenced *hph* gene, apparently by 'quelling' of *dim-5*. We demonstrate that recombinant DIM-5 protein specifically methylates histone H3 and that replacement of lysine 9 in histone H3 with either a leucine or an arginine phenocopies the *dim-5* mutation. We conclude that DNA methylation depends on histone methylation.

Cytosine methylation is essential for the normal development of mammals and plants. Mutations in any of three known DNA methyltransferase (DMTase) genes of the mouse (*Dnmt1*, *Dnmt3a* and *Dnmt3b*) are lethal, either during embryogenesis or soon thereafter^{1,2}. In humans, a syndrome characterized by immuno-deficiency, centromere instability and facial anomalies results from mutations in the *DNMT3B* gene³. Reduction in methylation in the plant *Arabidopsis thaliana*, caused either by expression of an anti-sense construct against the DMTase MET1, or by mutation of a gene encoding a putative chromatin remodelling factor (*ddm1*), results in developmental abnormalities and partial female sterility⁴.

In contrast to the situation in plants and animals, DNA methylation is not essential in the filamentous fungus *N. crassa*, thereby facilitating investigations of DNA methylation in this organism. As a step towards exploring the control and mechanism of cytosine methylation—processes that remain largely unknown in eukaryotes—we searched for methylation mutants in *Neurospora*. A screen of strains surviving a chemical mutagenesis yielded one mutant that was completely defective in methylation (*dim-2*), and another with an approximately 50% reduction in total DNA methylation (*dim-3*)⁵. The *dim-2* gene has recently been isolated and demonstrated to encode a DMTase that is responsible for both *de novo* and maintenance methylation at symmetrical and non-symmetrical sites⁶. Mutations in *dim-2* relieve silencing of methylated genes^{7,8} but do not noticeably affect growth or development⁶. Mutants with defects in an unidentified *Neurospora* gene affecting methylation, *dim-1*, were found among 5-azacytidine-resistant derivatives of a DNA repair mutant, *mus-20* (ref. 9). The current study arose from an attempt to tag the *dim-1* gene by insertional mutagenesis in a *mus-20* strain.

We unexpectedly generated a mutation in a previously unknown gene required for DNA methylation, *dim-5*, and mapped the mutation to an 80-kilobase (kb) region. This region included a gene homologous to histone methyltransferases, which is required for heterochromatin formation in fission yeast, *Drosophila* and mammals. Transformation experiments confirmed that the candidate gene is *dim-5*, and biochemical tests on recombinant DIM-5 demonstrated that this protein methylates histone H3. The implication that histone methylation controls DNA methylation was supported by demonstrating that replacements of K9 in histone H3 cause loss of DNA methylation *in vivo*.

Isolation and genetic mapping of the *dim-5* mutation

We selected 150 strains that were resistant to 5-azacytidine from

about 12,000 pRAL1 (*qa-2*⁺) transformants of an *aro-9 qa-2 mus-20* strain (N2141), and tested them for methylation defects by Southern hybridization using probes for the ζ - η (ref. 5) and Ψ 63 (ref. 10) methylated regions. One strain showed greatly reduced methylation in both regions. The methylation defect segregated in genetic crosses as expected of a normal mendelian allele; however, it did not co-segregate with either 5-azacytidine resistance or a pRAL1 insertion. Moreover, complementation tests between the new *dim* strain and previously identified mutants (*dim-1*, *dim-2* and *dim-3*) demonstrated that the mutant represents a new complementation group, which we designated *dim-5* (data not shown). Before detailed characterization, the *dim-5* mutation was purified from the mutagenized *mus-20* background by five backcrosses to a wild-type strain. Strains of *dim-5* showed slow, irregular growth (Fig. 1), unlike other methylation mutants, including *dim-2* (DMTase) null mutants (see Supplementary Information). In addition, homozygous *dim-5* crosses revealed a partial barren phenotype: few spores were produced, and most of those produced were inviable (data not shown).

About 1.5% of the cytosines in *N. crassa* DNA are methylated⁵ and

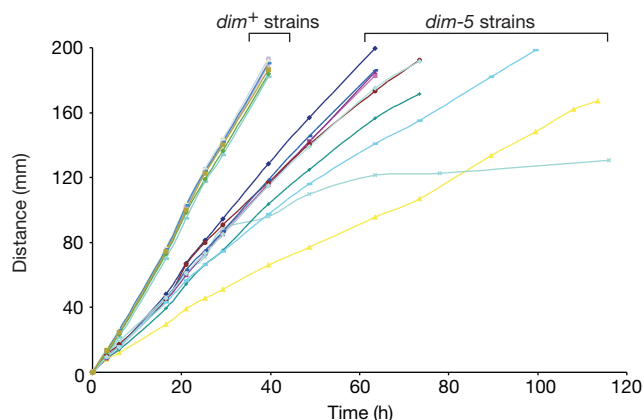


Figure 1 Growth deficiencies of *dim-5* strains. Rates of apical growth of ten wild-type (*dim*⁺) and ten mutant (*dim-5*) progeny of N2140 (*dim-5 leu-2 pan-2 A*) $\times$ N185 (*trp-4 a*) were measured at 32 °C using 'race tubes' containing 1.5% sucrose Vogel's medium with pantothenate. The average rates and standard deviations of the wild-type and mutant strains were 4.9 ± 0.1 and 2.4 ± 0.7 mm h⁻¹, respectively. The rates of the ten wild-type strains were so similar that their plots are virtually superimposed.

all known methylation is in relics of repeat-induced point mutation (RIP) and ribosomal DNA¹¹. We therefore analysed four relics of RIP (ζ - η ¹¹, Ψ 63 (ref. 10), 1D21 and 9A20; T. Wolfe and E.U.S., unpublished data) and rDNA for methylation in *dim-5* strains. Southern hybridizations using the isoschizomers *Dpn*II and *Sau*3AI, which both recognize GATC but differ in that only *Sau*3AI is inhibited by cytosine methylation, revealed no methylation in *dim-5* strains at any of the five regions tested (Fig. 2). Inspection of total genomic DNA digested with *Dpn*II or *Sau*3AI and stained with ethidium bromide indicated that *dim-5*, like *dim-2*, eliminates all, or nearly all, DNA methylation (Fig. 2).

To identify the *dim-5* gene, we first mapped it by conventional genetics, scoring the mutation by Southern hybridization. Results of initial crosses localized the gene to linkage group IV. Analysis of three-point linkage data from a cross between a *dim-5 pan-1 pyr-2* strain (N2142) and a *trp-4* strain (N185) placed *dim-5* approximately two map units in a centromere-distal position of *trp-4* (see Supplementary Information). To determine whether *dim-5* is in the 2–4-map-unit region between *trp-4* and *leu-2*, or in a centromere-distal position of *leu-2*, we tested methylation of DNA from *trp-4*⁺ *leu-2*⁺ recombinant progeny from a cross of the *trp-4* strain with a

dim-5 leu-2 strain (N2140). Fifteen out of 42 recombinants were defective in methylation, establishing that *dim-5* is between *trp-4* and *leu-2*, a region in which no mutation had been mapped previously¹².

Identification of the *dim-5* open reading frame

As a step towards identifying the *dim-5* gene, we searched *N. crassa* genomic sequence data (Neurospora Sequencing Project; <http://www-genome.wi.mit.edu>) for *leu-2* and *trp-4* on the basis of their expected homology to *Saccharomyces cerevisiae* *LEU1* and *TRP4*, respectively¹². Candidates for both genes were found separated by about 80 kb, consistent with the genetic distance between these genes. We examined the interval between the putative *leu-2* and *trp-4* genes for *dim-5* candidates (Fig. 3a). One of 15 candidates identified using BLASTx is predicted to encode a protein that is related to chromatin-associated proteins involved in gene silencing in fission yeast and fruit flies, namely *Schizosaccharomyces pombe* *Clr4* (ref. 13) and *Drosophila melanogaster* *Su(var)3-9* (ref. 14). Mutations in *clr4* derepress silent mating-type genes and genes inserted in other repressive domains and cause aberrant chromosome behaviour¹⁵, whereas mutations in *su(var)3-9* suppress the gene silencing phenomenon known as position effect variegation in *Drosophila*¹⁴. To test the possibility that the related *Neurospora* gene is required for DNA methylation in *Neurospora*, we tested fragments from the *leu-2/trp-4* region for complementation of the *dim-5* mutation. Because co-transformation is highly efficient in *Neurospora*, we transformed a *dim-5* strain with a mixture of a plasmid (pBT6) conferring resistance to benomyl (*Bml*^r) plus a test fragment, and then assayed random *Bml*^r transformants for *de novo* methylation of Ψ 63. Southern blot analysis of 26 *Bml*^r transformants generated with a 4.3-kb DNA fragment containing the *clr4/su(var)3-9* homologue, plus a putative 3-hydroxyisobutyrate dehydrogenase (*hibD*) gene, showed that all 24 transformants that incorporated the non-selected fragment displayed substantial *de novo* DNA methylation at the Ψ 63 and ζ - η regions and in the genome overall (see Supplementary Information). Equivalent results were obtained with smaller restriction fragments including the *clr4/su(var)3-9*-related gene alone (Fig. 3a, b). We therefore tentatively concluded that this open reading frame (ORF) is *dim-5*.

Derepression of a silenced transgene by quelling *dim-5*

It remained formally possible that ectopic insertions of the *Neurospora clr4/su(var)3-9* homologue suppressed the *dim-5* mutation owing to a dosage effect, but was not itself *dim-5*. We took advantage of quelling—a post-transcriptional gene silencing mechanism of *Neurospora* that does not depend on DNA methylation¹⁶—to carry out an independent test of the *clr4/su(var)3-9* homologue. In particular, we tested whether introduction of fragments of the presumptive *dim-5* coding region into a *Dim*⁺ *Neurospora* strain (N644) carrying a methylated, silent transgene (*hph*), would silence the wild-type *dim-5* gene and thereby activate the transgene (Fig. 4a). We co-transformed N644 with a mixture of pBT6 and an *Eco*RV digest of a PCR fragment containing the entire presumptive *dim-5* ORF, and tested *Bml*^r transformants for resistance to hygromycin (*Hyg*^r), which should result if the *hph* gene were demethylated. Five out of twelve *Bml*^r transformants that were generated with the DNA mixture showed strong resistance to hygromycin and several others showed partial resistance (Fig. 4b). In contrast, none of the control transformants generated with pBT6 alone were resistant to hygromycin. To test directly for loss of DNA methylation, we isolated DNA from these strains grown under selection for benomyl, but not hygromycin, and assayed methylation at the Ψ 63 region (Fig. 4c). Substantial hypomethylation was observed and was correlated with activation of the silenced *hph* gene. These results therefore confirmed our tentative identification of *dim-5* and demonstrated that this gene is subject to quelling.

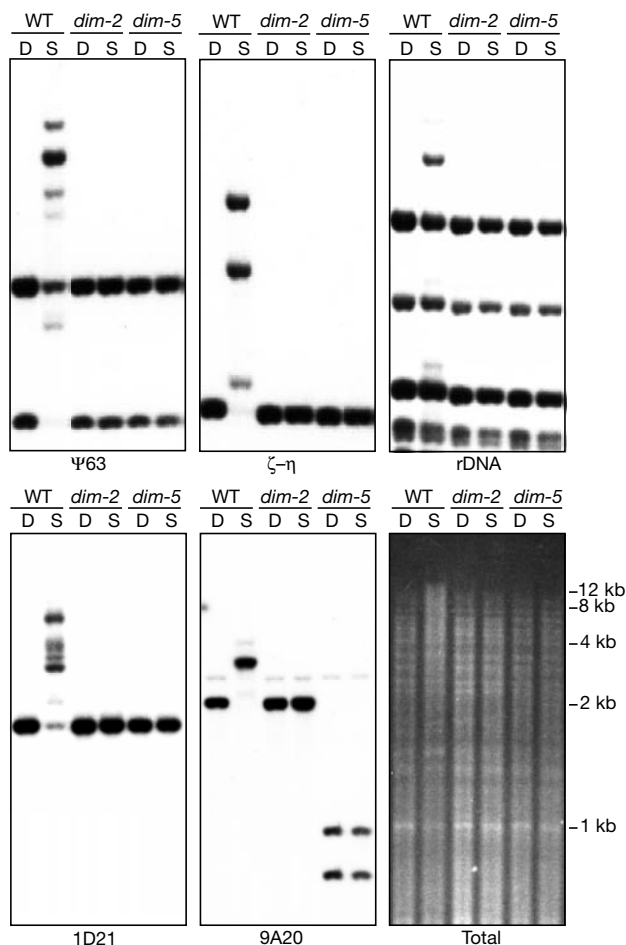


Figure 2 DNA methylation defect of *dim-5* strains. Genomic DNA of wild type (WT), a DNA methyltransferase mutant (*dim-2*) and the *dim-5* mutant were digested with *Dpn*II (D) or *Sau*3AI (S), and analysed by gel electrophoresis and Southern hybridization using probes for the indicated five methylated chromosomal regions. The DNAs were stained with ethidium bromide (Total) to reveal the total digestion profiles generated with these isoschizomers. Blots were reprobed for unmethylated regions to confirm that digests were complete. The 9A20 region in the *dim-5* strain shows a restriction-fragment length polymorphism relative to the wild-type and *dim-2* strains.

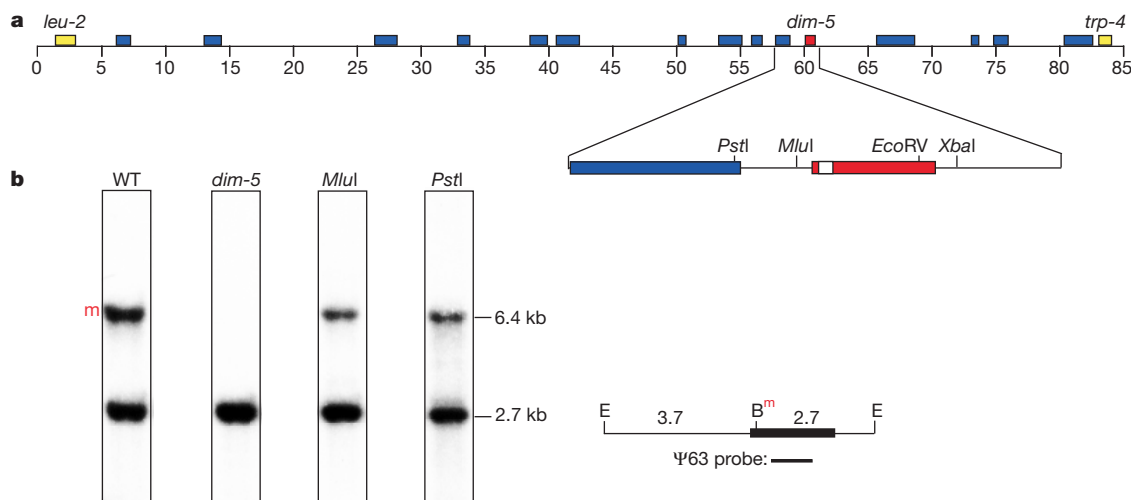


Figure 3 Identification of *dim-5* by genetic mapping and complementation. **a**, Map of genes revealed by BLASTx in *leu-2/trp-4* interval of contig 1.118 (assembly version 1; <http://www-genome.wi.mit.edu>). Regions of marked similarity (alignment scores >80) to genes in NCBI database are indicated (rectangles). A segment that was amplified to test two *dim-5* candidates (*hibD* gene, blue; homologue of *S. pombe* *clr4*, red with white intron) is shown expanded. **b**, Complementation of *dim-5* mutation. *dim-5* strain N2145

was co-transformed with pBT6 and a 2.0-kb *PstI/XbaI* or 1.4-kb *MluI/XbaI* fragment. Genomic DNA from representative Bm1^r transformants was analysed by Southern hybridization for DNA methylation in the Ψ63 region using *EcoRI* (E) and *BamHI* (B). Methylation of the *BamHI* site¹⁰ gives a 6.4-kb fragment, as shown (red m). Results for representative transformants are shown with positive (WT) and negative (*dim-5*) controls.

The *dim-5* mutant has a nonsense mutation in the SET domain

As a final test of whether we had correctly identified the *dim-5* gene, we amplified (using PCR) and sequenced the ORF from our mutant and from its wild-type parental strain. We found a single C to G mutation in the serine codon (TCA) at amino acid position 216 of the predicted 318-amino-acid polypeptide. The mutation generated a stop codon in the middle of a distinctive ~130-amino-acid sequence motif called the SET domain (Fig. 5). We conclude that DIM-5 is a SET domain protein, which is homologous to genes required for heterochromatin formation.

The SET domain was initially identified as a region of apparent homology in three nuclear proteins of *Drosophila*: Su(var)3-9, the Polycomb group protein E(Z) and Trithorax group protein TRX¹⁴. Greater than 200 genes with SET domains are now known¹⁷. Like *clr4*, *su(var)3-9*, *SUV39H1*, *Suv39h1* and *Suv39h2*, *dim-5* includes cysteine-rich sequences that flank a SET domain (Fig. 5). Noting that some SET proteins are protein methyltransferases, the possibility that chromatin-associated SET domain proteins are histone methyltransferases (HMTases) was examined¹⁸. Although not detected with all the chromatin-associated SET proteins, *S. pombe* Clr4 (refs 18, 19) and the closely related proteins from humans (*SUV39H1* (ref. 18) and mouse (*Suv39h1* (ref. 18) and *Suv39h2* (ref. 20) did show HMTase activity. Analyses of variants generated *in vitro* indicated that both the SET domain and the associated cysteine-rich sequences are required for HMTase activity^{18,19}.

dim-5 encodes a histone H3 methyltransferase

We directly investigated the possibility that DIM-5, like Clr4 and Suv39h2, is a histone methyltransferase. We built an inducible glutathione S-transferase (GST) fusion construct containing nearly all of the DIM-5 coding region, similar to constructs made to assay HMTase activity of Clr4 and Suv39h2 (Fig. 5b). Recombinant DIM-5 fusion protein was purified from *Escherichia coli*, provided with S-adenosyl-[methyl-3H]-L-methionine as a potential methyl-group donor, and incubated with a natural mixture of histones from calf thymus. The proteins were then fractionated by SDS-polyacrylamide gel electrophoresis (PAGE) and assayed for incorporation of methyl groups by fluorography and scintillation counting of gel slices. We detected significant incorporation of

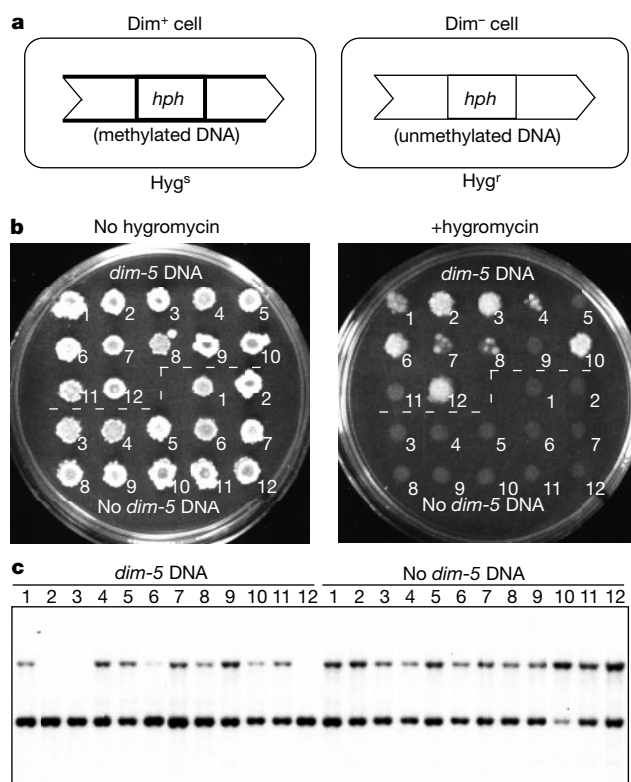


Figure 4 Quelling of *dim-5* relieves silencing of *hph* in N644. **a**, Sketch of methylated (*Hyg*^R) or unmethylated (*Hyg*^S) *hph* gene, flanked by methylation-inducing DNA segments that had been subjected to RIP⁴⁹. **b**, Conidia of 24 random transformants of strain N644, generated with pBT6 alone or with pBT6 plus a PCR fragment digested with *EcoRV* containing *dim-5* (see Methods), were spot-tested for growth in the absence or presence of hygromycin. **c**, Effect of *dim-5* gene fragments on methylation at Ψ63. The transformants shown in **b** were analysed for methylation as in Fig. 3.

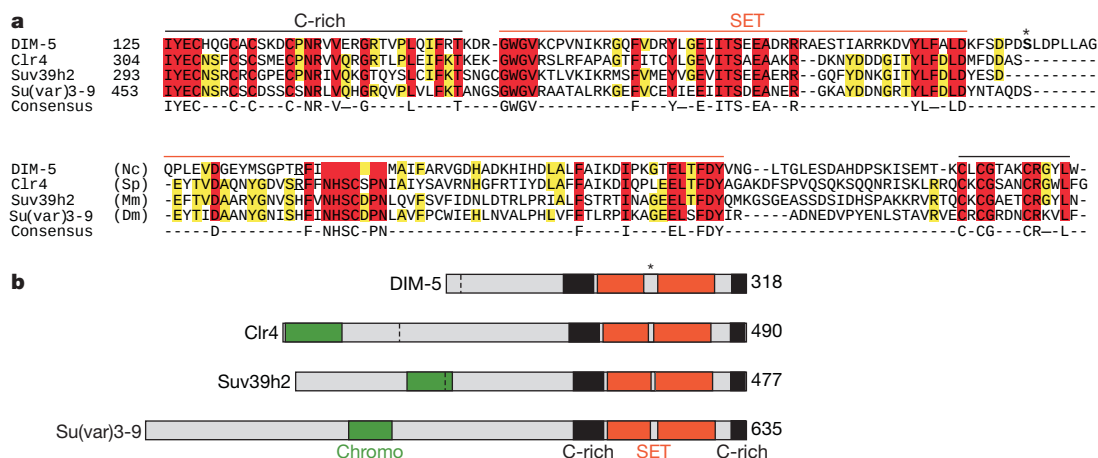


Figure 5 DIM-5 is a SET-domain protein. **a**, Amino acid alignment of conserved regions of *N. crassa* DIM-5 (accession number AF419248), *S. pombe* Clr4 (accession number O60016), *Mus musculus* Suv39h2 (accession number NM 022724.1) and *D. melanogaster* Su(var)3-9 (accession number S47004) generated with CLUSTALW (<http://workbench.sdsc.edu>). Amino acid coordinates and species abbreviations are shown at the left. Nc, *N. crassa*; Sp, *S. pombe*; Mm, *M. musculus*; Dm, *D. melanogaster*. The position of a nonsense mutation in allele *dim-5* HT1 is indicated with an asterisk. Amino acid residues conserved among all four proteins are highlighted in red and

indicated in 'consensus'; residues conserved in three of the four proteins are highlighted in yellow; dashes indicate gaps in the alignment. The black and orange lines mark the extent of the cysteine-rich (C-rich) and SET domains, respectively. **b**, Protein domain organization of DIM-5 and related proteins aligned at their carboxy termini with predicted number of amino acids and locations of chromo (green) SET (orange) and cysteine-rich (black) domains indicated. The amino-terminal endpoints of recombinant proteins made in this study or previously¹⁸ are indicated by vertical dashed lines.

labelled methyl groups into histones, indicating that DIM-5 is an HMTase. Moreover, histone H3 was the acceptor of nearly all the methylation, as found with Clr4 and Suv39h2 (Fig. 6). Much weaker incorporation of label was detected in a protein tentatively identified as a histone H1 (data not shown).

Lysine 9 of histone H3 gene is involved in DNA methylation

Our demonstration that DIM-5 has HMTase activity suggested that

methylation of histone H3 is necessary for DNA methylation in *Neurospora*. It remained possible, however, that the HMTase activity was not relevant to its role in DNA methylation. We therefore sought *in vivo* evidence for involvement of histone H3 in DNA methylation. The Clr4 and Suv39h2 HMTases are specific for K9 of H3 (refs 18, 19), a residue that can be either methylated or acetylated¹⁷ (Fig. 7a). We reasoned that if methylation of K9 of histone H3 were required for DNA methylation in *Neurospora*, then replacement of this residue with an amino acid that is not subject to this modification should interfere with DNA methylation. We therefore mutated the H3 gene (*hH3*) *in vitro*, replacing the lysine codon with codons for leucine or arginine, and introduced the modified genes into strain N644 by co-transformation. Leucine and arginine were chosen for the following reasons: (1) they are structurally similar to lysine; (2) the neutral amino acid leucine can be regarded as a mimic of an acetylated lysine; (3) the positively charged amino acid arginine can be regarded as a mimic of an unacetylated lysine; and (4) leucine is known not to be a substrate for methylation of recombinant Suv39h1 HMTase¹⁸.

Neurospora crassa has only a single copy of the *hH3* gene²¹ but gene replacement by homologous recombination is inefficient in *Neurospora*. Furthermore, replacement of the wild-type *hH3* with the mutated versions might be lethal. It seemed possible, however, that the mutations would prove dominant or semi-dominant. We therefore took advantage of the methylated *hph* allele of N644 to test for loss of DNA methylation in random transformants generated with the mutated *hH3* constructs. Transformants generated with mutant or wild-type *hH3* genes together with the co-transformation marker, *Bml*, were selected *en masse* on benomyl medium and then tested for expression of *hph*. About 500 asexual spores from each pool, representing roughly 30 *Bml*^r transformants, were spread on hygromycin plates (Fig. 7b). Hyg^r colonies were obtained with both mutants, but not with the wild-type control, in two independent experiments. Southern hybridization analyses of DNA isolated from representative Hyg^r transformants demonstrated that the strains had markedly reduced DNA methylation at Ψ 63, and confirmed that they contained modified *hH3* alleles (Fig. 7c). The Hyg^r transformants in each experiment contained a single ectopic copy of the mutant allele, which is unusual for *Neurospora*. Perhaps

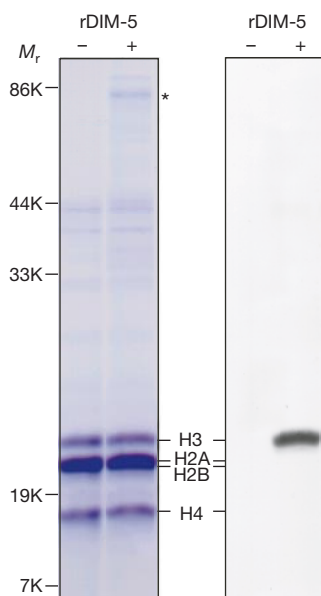


Figure 6 Histone methyltransferase activity of recombinant DIM-5 protein (rDIM-5). Purified histones (~20 μ g; Boehringer Mannheim) were incubated for 6 h at 20 °C with or without purified GST-DIM-5 fusion protein (GST-DIM-5; ~1 μ g) and 2.75 μ Ci S-adenosyl-[methyl-3H]-L-methionine, as methyl donor. Reaction products were fractionated by PAGE (16.5%), stained with Coomassie blue (left) and then fluorographed (right) to detect methylation. The positions of selected size standards, intact recombinant protein (asterisk) and core histones are indicated.

additional copies of the mutant *hH3* genes were toxic, either directly or because they caused quelling, thereby reducing H3 levels beyond the point that the cells could survive. Direct DNA sequencing of *hH3* PCR products confirmed that the strains contained both the wild-type and mutant sequences (Fig. 7c). These results strongly support the inference from our other results that methylation of histone H3 is critical for DNA methylation.

Does RIP generate heterochromatin?

The densely staining 'constitutive heterochromatin', which is frequently found near centromeres and telomeres in eukaryotes, provides an example of how DNA sequences can direct the formation of specialized forms of chromatin. Constitutive heterochromatin is typically rich in moderately repeated sequences (such as transposons) and highly repeated sequences (such as satellite DNA) and displays a number of other identifying characteristics. It remains condensed after mitosis, replicates late in S phase, shows low levels of genetic recombination, contains special forms of histones, and, in organisms with DNA methylation such as mammals and plants, it is hypermethylated²². Classical studies in *Drosophila* have demonstrated that heterochromatin is a repressive environment for most genes; rearrangements that move heterochromatic regions typically cause silencing of nearby genes in some fraction of the cells. Isolation of suppressors of the resulting variegation identified the heterochromatin protein HP1 (Su(var)3-5)²³, an HMTase (Su(var)3-9)^{14,18} and an S-adenosyl-methionine synthetase (Su(z)5)²⁴. The very much smaller heterochromatic regions of *S. pombe* depend on a similar set of silencing genes, including the chromo domain genes *swi6* and *clr4*, which encode a HP1-like protein and a histone H3 MTase, respectively¹⁹. The chromo domain of HP1 has recently been shown to recognize methylated K9 of histone H3 (refs 25, 26).

Our finding that DNA methylation in *Neurospora* relies on histone H3 K9 raises the possibility that sequences mutated by RIP, which constitute the bulk of methylated sequences of this organism and are found concentrated in centromeric DNA²⁷, serve to nucleate heterochromatin. RIP detects duplicated sequences, such as transposons^{10,27}, during the sexual phase of the *Neurospora*

life cycle and peppers them with G-C to A-T mutations¹¹. Like some satellite sequences, products of RIP are bound by proteins that show limited sequence specificity (H.T., G. Kothe and E.U.S., unpublished observations). These proteins may be responsible for recruiting heterochromatin proteins—perhaps the DIM-5 HMTase. Thus, relics of RIP may underlie heterochromatin in *Neurospora* and serve the cell for centromere function. Our observation that *dim-5* strains grow irregularly and are nearly sterile in homozygous crosses, unlike other known DNA methylation mutants, indicates that DIM-5 is involved in process(es) besides DNA methylation such as in heterochromatin formation.

Histones as signal transducers for DNA methylation

The control of DNA methylation has remained unclear despite decades of intensive investigations in mammals, plants and fungi. Although prokaryotic and eukaryotic DMTases show marked structural similarities, prokaryotes offer an inappropriate model for DNA methylation in eukaryotes. Bacterial DMTases require nothing more than DNA and a methyl-group donor for proper function, and are sequence specific. In contrast, eukaryotic DMTases have substantial non-catalytic domains that reflect interactions with other proteins²⁸, and show little sequence specificity²⁹. We suggest that eukaryotic DMTases evolved to take their cues primarily from chromatin. A common view is that histones act as obstacles to proteins that need access to the DNA. The discovery of chromatin-remodelling factors that can move nucleosomes³⁰, and of a putative chromatin-remodelling factor involved in DNA methylation³¹, fit this view. Our discovery that DNA methylation depends on histone methylation in *Neurospora* suggests an alternative possibility, however. Proteins that require access to DNA, such as DMTases, may actually depend on histones as cofactors. According to this view, chromatin-remodelling factors may assist histone modification enzymes to reach their targets and/or to facilitate exchange of nucleosomes with different modification states.

Histones are well suited to integrate information relevant to whether DNA in a particular region should be methylated. In addition to their proximity to DNA, histones are subject to a variety

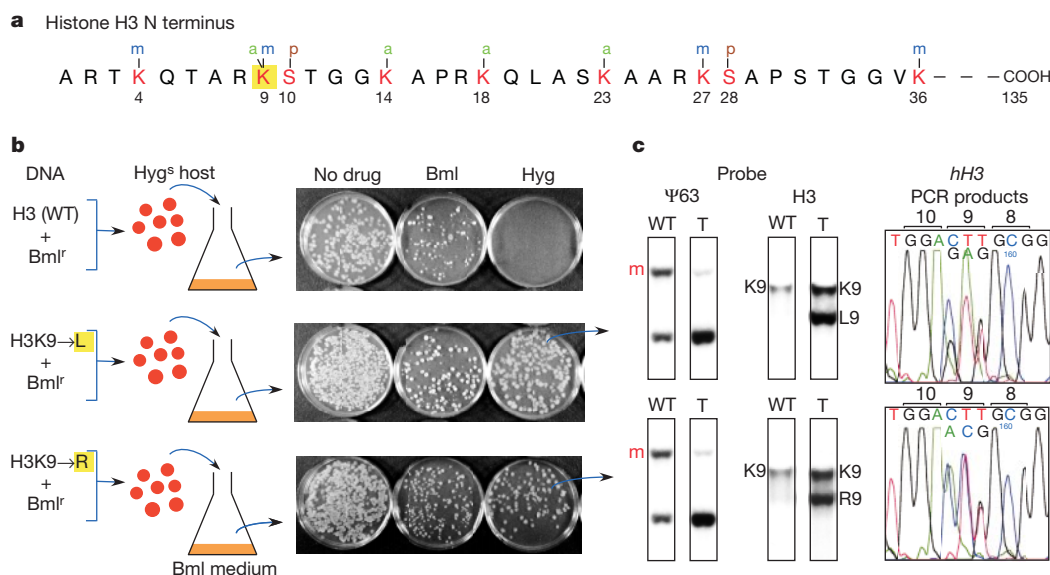


Figure 7 Reactivation of *hph* and loss of DNA methylation induced by transformation of a *dim*⁺ strain with mutant alleles of histone H3 gene (*hH3*). **a**, Sequence of the N-terminal segment of *Neurospora* histone H3 with residues presumed to be subject to methylation (m), acetylation (a) or phosphorylation (p) indicated in red and the residue implicated in silencing highlighted in yellow. **b**, Transformation experiment. Strain N644 (see Fig. 4) was co-transformed with pBT6 and wild-type or mutant versions of *hH3*, and Bml^r

transformants were selected *en masse* on benomyl and tested for drug resistance. **c**, Southern analysis and sequencing of DNA from Hyg^r transformants. DNA of representative transformants (T) and a wild-type (WT) control grown non-selectively was analysed with *Eco*RI and *Bam*HI for methylation (m) at Ψ63 as in Fig. 3, and for ectopic alleles of *hH3*. Direct sequencing of *hH3* PCR products confirmed the presence of both the wild-type and mutant alleles in representative strains (sequencing chromatograms).

of post-translational modifications (phosphorylation, methylation, acetylation, ubiquitination and ADP-ribosylation) that can have informational roles in the cell³². Acetylation, currently the best understood modification, is controlled by histone acetylases (HATs) and histone deacetylases (HDACs), which typically act as transcriptional coactivators and corepressors, respectively. Mechanistic connections are emerging among DNA methylation, histone deacetylation and histone methylation. Mammalian methyl-DNA binding proteins and DMTases associate with HDACs, and the histone deacetylase inhibitor trichostatin A (TSA) causes selective loss of DNA methylation in *Neurospora*, suggesting that histone acetylation can influence DNA methylation^{33,34}. Moreover, in *S. pombe*, TSA treatment or mutation of HDAC genes causes mislocalization of Swi6 and other defects characteristic of disruption of the *clr4* HMTase^{19,35,36}, perhaps because methylation of K9 of histone H3 is inhibited by acetylation of K9 or 14 (refs 18, 19). Phosphorylation of S10 also strongly inhibits methylation of K9 (ref. 18), providing another illustration of how histones can integrate information from multiple inputs and act as signal transducers. There are potentially a huge number of combinations of variously modified histones, and the network that connects modifications of DNA and chromatin proteins may be complex.

De novo and maintenance methylation of DNA in eukaryotes

A defining feature of epigenetic states is that they promote their own propagation. Thus, active chromosomal regions are rarely silenced and silenced regions are rarely activated. It was recognized previously that the symmetry of methylated sites (5'-CpG/GpC-5') in mammalian DNA would support a simple mechanism to propagate methylation patterns; all that was required was a DMTase specific for hemi-methylated sites³⁷. The 'maintenance methylase' model was supported by evidence that methylation states are indeed propagated, and by the discovery of DMTases that prefer hemi-methylated substrates. Nevertheless, the classic maintenance model does not account for some observations, such as heterogeneous methylation in cell clones, spreading of methylation and stable propagation of methylation at non-symmetrical sites, as observed in *Neurospora* and other eukaryotes^{38–40,50}. Now that we know that histone modifications can impact both *de novo* and maintenance DNA methylation, we should consider the possibility that propagation of DNA methylation patterns in eukaryotes depends on feedback loops between modifications of chromatin proteins and DNA. Of note, histones H3 and H4 remain tightly bound to DNA *in vivo*, unlike histones H2A and H2B⁴¹. This is consistent with the idea that these histones are involved in the propagation of epigenetic states.

In *Drosophila* and *S. pombe*, interplay between heterochromatin proteins propagates epigenetic states without relying on DNA methylation^{42–44}. If the chromo domains of Clr4 and Su(var)3-9 recognize methylated K9 of histone H3, as with HP1 (refs 25, 26), this might lead to preferential methylation of histones in previously methylated regions, which would propagate silencing. The absence of the chromo domain from the DIM-5 HMTase, and from the recently described G9a HMTase⁴⁵, is interesting. Perhaps DNA methylation and associated factors (for example, DMTases and methyl-DNA binding proteins) substitute for this potential self-reinforcing system. DMTases containing a chromo domain have been identified in plants⁴⁶, suggesting that some DMTases may take cues directly from histones. A search of public databases with DIM-5 revealed a number of potential HMTases that may be involved in DNA methylation. Future work will reveal the detailed relationship between methylation of histone H3 and DNA in *Neurospora*, and will show whether this 'double methylation' silencing system is widely used in eukaryotes. □

Methods

PCR primers are listed in Supplementary Information.

Analysis of DNA methylation

Genomic DNA was isolated from liquid cultures grown for 2 days at 32 °C, and analysed for DNA methylation by Southern hybridization as previously described³. The probe for the Ψ 63 region was a 0.9-kb *Ban*II/*Eco*O109 fragment isolated from pPG22 (ref. 10). The 0.8-kb *Bam*HI fragment was used to probe for the ζ - η region, and a 9.2-kb *Kpn*I fragment representing one repeat unit of the rDNA was used to probe for rDNA. The 1D21 and 9A20 probes were generated by PCR from the wild-type strain 74-OR23-IVA. Strains used for the methylation analysis shown in Fig. 2 were: 74-OR23-IVA (wild type; all blots), *dim-2* strains N1275 (ref. 5) (*dim-2 arg-10 mat A*; Ψ 63, ζ - η and rDNA blots) and N1877 (ref. 6) (*dim-2::hph his-3 mat a*; 1D21 and 9A20 blots), and *dim-5* strains N2144 (*dim-5*; Ψ 63, ζ - η and rDNA blots), N2145 (*dim-5 leu-2 pan-2 a*; 1D21 blot) and N2140 (*dim-5 leu-2 pan-2 A*; 9A20 blot).

Complementation tests and quelling experiments

Complementation of *dim-5* was accomplished with either a 4.3-kb *hibD/dim-5* PCR fragment amplified from wild-type strain 74-OR8-1a with *Pfu* DNA polymerase (Promega), or restriction fragments of this PCR fragment (Fig. 3a). Strain N2145 was co-transformed by electroporation⁴⁷ using 100 μ g pBT6, which confers resistance to benomyl (ref. 48), and 300 μ g test DNA.

For quelling, a PCR fragment containing the presumed *dim-5* ORF was generated, cut within the SET domain of the gene with *Eco*RV, and introduced into strain N644 (ref. 49) (*dim-5⁺ am^{RIP}/hph/am^{RIP} am¹³² inl a*) by co-transformation with pBT6. Approximately 10⁴ conidia of representative Bml^r transformants were spot-tested on Vogel's sorbose medium in the presence or absence of hygromycin (200 μ g ml⁻¹; Calbiochem). Genomic DNA was isolated from liquid cultures grown for 2 days at 32 °C, and analysed for DNA methylation⁵.

Identification of mutation in *dim-5* allele HT1

To identify the mutation in *dim-5* allele HT1 of strain N2140, pooled PCR products from six independent reactions produced with *Pfu* DNA polymerase and the *dim-5* ORF primers were gel-purified and sequenced directly on both strands. The wild-type allele (Fungal Genetics Stock Center (FGSC) 988) was isolated and sequenced in the same way (GenBank accession code AF419248).

Generation and purification of GST-DIM-5 fusion protein

A segment of the *dim-5* ORF, including amino acid residues 19–318, was amplified from *Neurospora* wild-type strain 74-OR8-1a (FGSC 988) with *Pfu* DNA polymerase and GST-DIM-5 primers. The PCR product was digested with *Bam*HI and *Eco*RI, gel-purified and cloned into the GST fusion expression vector pGEX-5X-3 (Pharmacia) using *E. coli* strain DH5 α F'. Recombinant protein was prepared from an 800-ml culture of *E. coli* cells grown 3 h at 37 °C in LB medium with ampicillin (400 μ g ml⁻¹), shifted to 30 °C for 40 min, induced with IPTG (0.1 mM) and collected 1 h later. Cells were lysed by sonication on ice in 8 ml of RIPA buffer (20 mM Tris (pH 7.5), 500 mM NaCl, 5 mM EDTA, 1% IGEPAL CA-630 (Sigma), 0.5% deoxycholate)¹⁸ with 5 mg ml⁻¹ lysozyme and a proteinase inhibitor cocktail (Complete; Boehringer Mannheim). The extract was clarified by centrifugation and the soluble proteins were incubated at 4 °C for 15 min with 700 μ l glutathione-agarose (Sigma) equilibrated in RIPA. GST/DIM-5 protein was purified by washing three times with 30 ml RIPA buffer, followed by elution from glutathione-agarose using 75 mM HEPES (pH 7.9), 150 mM NaCl, 5 mM dithiothreitol and 10 mM reduced glutathione (Sigma). The eluate was concentrated with a Centricon-30 filter (Millipore).

Histone methyltransferase assay

HMTase assays were carried out on a natural mixture of calf thymus histones as described¹⁸, except that the reaction was carried out at 20 °C for 6 h with 2.75 μ Ci S-adenosyl-[methyl-3H]-L-methionine (0.55 mCi ml⁻¹; NEN). Products were fractionated on SDS-PAGE (16.5%; 29:1) gels and fluorographed (4–12 h) using ENTENSIFY (DuPont).

In vitro mutagenesis and in vivo analysis of histone H3 variants

K9 of *N. crassa* histone H3 (ref. 21) was changed to leucine and arginine using the PCR-based QuickChange site-directed mutagenesis protocol (Stratagene) with a 4.9-kb plasmid carrying the wild-type H3 gene (*hH3*) and 1,161 bp of 5'-flanking sequences (pSH12; S. Hays and E.U.S., unpublished data) as template. Primer pairs H3L9 and H3R9 were used to generate CTC and CGT codons in place of the AAG codon, respectively. The resulting plasmids (pSH12L9 and pSH12R9, respectively) and the wild-type control were linearized using *Xba*I and co-transformed into *N. crassa* strain N644 along with pBT6 that was linearized with *Hind*III. Approximately 1,000 conidia from transformants grown *en masse* on solidified Vogel's sucrose medium in flasks were plated on media containing no drug, benomyl (0.5 μ g ml⁻¹) or hygromycin (200 μ g ml⁻¹). Random Hyg^r transformants, which were obtained with pSH12L9 and pSH12R9 but not with pSH12 (Fig. 7), were grown in liquid medium in the absence of hygromycin to isolate DNA. The presence of ectopic *hH3* sequences with the expected mutations was verified by Southern hybridization and direct sequencing of PCR products generated using H3-ORF primers.

Received 24 July; accepted 12 September 2001.

- Li, E., Bestor, T. H. & Jaenisch, R. Targeted mutation of the DNA methyltransferase gene results in embryonic lethality. *Cell* **69**, 915–926 (1992).
- Okano, M., Bell, D. W., Haber, D. A. & Li, E. DNA methyltransferases Dnmt3a and Dnmt3b are essential for *de novo* methylation and mammalian development. *Cell* **99**, 247–257 (1999).

3. Xu, G. L. *et al.* Chromosome instability and immunodeficiency syndrome caused by mutations in a DNA methyltransferase gene. *Nature* **402**, 187–191 (1999).
4. Mittelsten Scheid, O. & Paszkowski, J. Transcriptional gene silencing mutants. *Plant Mol. Biol.* **43**, 235–241 (2000).
5. Foss, H. M., Roberts, C. J., Claeys, K. M. & Selker, E. U. Abnormal chromosome behavior in *Neurospora* mutants defective in DNA methylation. *Science* **262**, 1737–1741 (1993).
6. Kouzminova, E. A. & Selker, E. U. *Dim-2* encodes a DNA-methyltransferase responsible for all known cytosine methylation in *Neurospora*. *EMBO J.* **20**, 4309–4323 (2001).
7. Rountree, M. R. & Selker, E. U. DNA methylation inhibits elongation but not initiation of transcription in *Neurospora crassa*. *Genes Dev.* **11**, 2383–2395 (1997).
8. Cambareri, E. B., Foss, H. M., Rountree, M. R., Selker, E. U. & Kinsey, J. A. Epigenetic control of a transposon-inactivated gene in *Neurospora* is dependent on DNA methylation. *Genetics* **143**, 137–146 (1996).
9. Foss, H. M., Roberts, C. J. & Selker, E. U. Reduced levels and altered patterns of DNA methylation caused by mutations in *Neurospora crassa*. *Mol. Gen. Genet.* **259**, 60–71 (1998).
10. Margolin, B. S. *et al.* A methylated *Neurospora* 5S rRNA pseudogene contains a transposable element inactivated by RIP. *Genetics* **149**, 1787–1797 (1998).
11. Selker, E. U. Premiotic instability of repeated sequences in *Neurospora crassa*. *Annu. Rev. Genet.* **24**, 579–613 (1990).
12. Perkins, D. D., Radford, A. & Sachs, M. S. *The Neurospora Compendium; Chromosomal Loci* (Academic, San Diego, 2001).
13. Ivanova, A. V., Bonaduce, M. J., Ivanov, S. V. & Klar, A. J. The chromo and SET domains of the Ctr4 protein are essential for silencing in fission yeast. *Nature Genet.* **19**, 192–195 (1998).
14. Tschiersch, B. *et al.* The protein encoded by the *Drosophila* position-effect variegation suppressor gene *Su(var)3-9* combines domains of antagonistic regulators of homeotic gene complexes. *EMBO J.* **13**, 3822–3831 (1994).
15. Allshire, R. C., Nimmo, E. R., Ekwall, K., Javerzat, J. P. & Cranston, G. Mutations derepressing silent centromeric domains in fission yeast disrupt chromosome segregation. *Genes Dev.* **9**, 218–233 (1995).
16. Cogoni, C. *et al.* Transgene silencing of the *al-1* gene in vegetative cells of *Neurospora* is mediated by a cytoplasmic effector and does not depend on DNA–DNA interactions or DNA methylation. *EMBO J.* **15**, 3153–3163 (1996).
17. Jenuwein, T. Re-SET-ting heterochromatin by histone methyltransferases. *Trends Cell Biol.* **11**, 266–273 (2001).
18. Rea, S. *et al.* Regulation of chromatin structure by site-specific histone H3 methyltransferases. *Nature* **406**, 593–599 (2000).
19. Nakayama, J., Rice, J. C., Strahl, B. D., Allis, C. D. & Grewal, S. I. Role of histone H3 lysine 9 methylation in epigenetic control of heterochromatin assembly. *Science* **292**, 110–113 (2001).
20. O'Carroll, D. *et al.* Isolation and characterization of *Suv39h2*, a second histone H3 methyltransferase gene that displays testis-specific expression. *Mol. Cell. Biol.* **20**, 9423–9433 (2000).
21. Woudt, L. P. *et al.* The genes coding for histone H3 and H4 in *Neurospora crassa* are unique and contain intervening sequences. *Nucleic Acids Res.* **11**, 5347–5361 (1983).
22. Hennig, W. Heterochromatin. *Chromosoma* **108**, 1–9 (1999).
23. Eissenberg, J. C. *et al.* Mutation in a heterochromatin-specific chromosomal protein is associated with suppression of position-effect variegation in *Drosophila melanogaster*. *Proc. Natl Acad. Sci. USA* **87**, 9923–9927 (1990).
24. Larsson, J., Zhang, J. & Rasmuson-Lestander, A. Mutations in the *Drosophila melanogaster* gene encoding S-adenosylmethionine suppress position-effect variegation. *Genetics* **143**, 887–896 (1996).
25. Bannister, A. J. *et al.* Selective recognition of methylated lysine 9 on histone H3 by the HP1 chromo domain. *Nature* **410**, 120–124 (2001).
26. Lachner, M., O'Carroll, D., Rea, S., Mechtler, K. & Jenuwein, T. Methylation of histone H3 lysine 9 creates a binding site for HP1 proteins. *Nature* **410**, 116–120 (2001).
27. Cambareri, E. B., Aisner, R. & Carbon, J. Structure of the chromosome VII centromere region in *Neurospora crassa*: degenerate transposons and simple repeats. *Mol. Cell. Biol.* **18**, 5465–5477 (1998).
28. Colot, V. & Rossignol, J. L. Eukaryotic DNA methylation as an evolutionary device. *Bioessays* **21**, 402–411 (1999).
29. Yoder, J. A., Soman, N. S., Verdine, G. L. & Bestor, T. H. DNA (cytosine-5)-methyltransferases in mouse cells and tissues. Studies with a mechanism-based probe. *J. Mol. Biol.* **270**, 385–395 (1997).
30. Kingston, R. E. & Narlikar, G. J. ATP-dependent remodeling and acetylation as regulators of chromatin fluidity. *Genes Dev.* **13**, 2339–2352 (1999).
31. Jeddeloh, J. A., Stokes, T. L. & Richards, E. J. Maintenance of genomic methylation requires a SWI2/SNF2-like protein. *Nature Genet.* **22**, 94–97 (1999).
32. Strahl, B. D. & Allis, C. D. The language of covalent histone modifications. *Nature* **403**, 41–45 (2000).
33. Dobosy, J. R. & Selker, E. U. Emerging connections between DNA methylation and histone acetylation. *Cell. Mol. Life Sci.* **58**, 721–727 (2001).
34. Selker, E. U. Trichostatin A causes selective loss of DNA methylation in *Neurospora*. *Proc. Natl Acad. Sci. USA* **95**, 9430–9435 (1998).
35. Grewal, S. I., Bonaduce, M. J. & Klar, A. J. Histone deacetylase homologs regulate epigenetic inheritance of transcriptional silencing and chromosome segregation in fission yeast. *Genetics* **150**, 563–576 (1998).
36. Ekwall, K., Olsson, T., Turner, B. M., Cranston, G. & Allshire, R. C. Transient inhibition of histone deacetylation alters the structural and functional imprint at fission yeast centromeres. *Cell* **91**, 1021–1032 (1997).
37. Bestor, T. H. & Tycko, B. Creation of genomic methylation patterns. *Nature Genet.* **12**, 363–367 (1996).
38. Singer, M. J., Marcotte, B. A. & Selker, E. U. DNA methylation associated with repeat-induced point mutation in *Neurospora crassa*. *Mol. Cell. Biol.* **15**, 5586–5597 (1995).
39. Miao, V. P., Freitag, M. & Selker, E. U. Short TpA-rich segments of the zeta-eta region induce DNA methylation in *Neurospora crassa*. *J. Mol. Biol.* **300**, 249–273 (2000).
40. Selker, E. U., Fritz, D. Y. & Singer, M. J. Dense non-symmetrical DNA methylation resulting from repeat-induced point mutation (RIP) in *Neurospora*. *Science* **262**, 1724–1728 (1993).
41. Kimura, H. & Cook, P. R. Kinetics of core histones in living human cells. Little exchange of h3 and h4 and some rapid exchange of h2b. *J. Cell Biol.* **153**, 1341–1354 (2001).
42. Lewis, E. B. The phenomenon of position effect. *Adv. Genet.* **3**, 73–115 (1950).
43. Allshire, R. C., Javerzat, J. P., Redhead, N. J. & Cranston, G. Position effect variegation at fission yeast centromeres. *Cell* **76**, 157–169 (1994).
44. Grewal, S. I. & Klar, A. J. Chromosomal inheritance of epigenetic states in fission yeast during mitosis and meiosis. *Cell* **86**, 95–101 (1996).
45. Tachibana, M., Sugimoto, K., Fukushima, T. & Shinkai, Y. SET domain-containing protein, G9a, is a novel lysine-preferring mammalian histone methyltransferase with hyperactivity and specific selectivity to lysines 9 and 27 of histone H3. *J. Biol. Chem.* **276**, 25309–25317 (2001).
46. Lindroth, A. M. *et al.* Requirement of CHROMOMETHYLASE3 for maintenance of CpXpG methylation. *Science* **292**, 2077–2080 (2001).
47. Margolin, B. S., Freitag, M. & Selker, E. U. Improved plasmids for gene targeting at the *his-3* locus of *Neurospora crassa* by electroporation. *Fungal Gen. Newsl.* **44**, 34–36 (1997).
48. Orbach, M. J., Porro, E. B. & Yanofsky, C. Cloning and characterization of the gene for β -tubulin from a benomyl-resistant mutant of *Neurospora crassa* and its use as a dominant selectable marker. *Mol. Cell. Biol.* **6**, 2452–2461 (1986).
49. Ireland, J. T. & Selker, E. U. Cytosine methylation associated with repeat-induced point mutation causes epigenetic gene silencing in *Neurospora crassa*. *Genetics* **146**, 509–523 (1997).
50. Ramsahoye, B. H. *et al.* Non-CpG methylation is prevalent in embryonic stem cells and may be mediated by DNA methyltransferase 3a. *Proc. Natl Acad. Sci. USA* **97**, 5237–5242 (2000).

Supplementary Information accompanies the paper on Nature's website (<http://www.nature.com>).

Acknowledgements

We are grateful to H. Inoue for a gift of *Neurospora* strain i47. We thank S. Hays for providing the plasmid containing wild-type hH3, and J. Dobosy, M. Freitag, S. Hays, G. Kothe, E. Kouzminova, J. Selker, T. Stevens and P. von Hippel for discussions and comments on the manuscript, and R. Metzner, D. Perkins and C. Yanofsky for encouragement. H.T. thanks S. Crosthwaite for inspiration. This study was supported by a grant from the NIH to E.U.S.

Correspondence and requests for materials should be addressed to E.U.S. (e-mail: selker@molbio.uoregon.edu). The GenBank accession code for wild-type and mutant *dim-5* sequences is AF419248.

Water-maser emission from a planetary nebula with a magnetized torus

L. F. Miranda*, Y. Gómez†, G. Anglada*‡ & J. M. Torrelles§

* Instituto de Astrofísica de Andalucía, CSIC, Apdo. Correos 3004, E-18080 Granada, Spain

† Instituto de Astronomía UNAM, Campus Morelia, A.P. 3-72 (Xangari), c.p. 58089 Morelia, Michoacán, Mexico

‡ Harvard-Smithsonian Center for Astrophysics, 60 Garden Street, MS-42, Cambridge, Massachusetts 02138, USA

§ Institut d'Estudis Espacials de Catalunya (IEEC/CSIC) and Institut de Ciències de l'Espai (CSIC), Edifici Nexus, C/Gran Capità 2-4, E-08034 Barcelona, Spain

A star like the Sun becomes a planetary nebula towards the end of its life, when the envelope ejected during the earlier giant phase becomes photoionized as the surface of the remnant star reaches a temperature of $\sim 30,000$ K. The spherical symmetry of the giant phase is lost in the transition to a planetary nebula, when non-spherical shells and powerful jets develop. Molecules that were present in the giant envelope are progressively destroyed by the radiation¹. The water-vapour masers that are typical of the giant envelopes^{2,3} therefore are not expected to persist in planetary nebulae^{1,4}. Here we report the detection of water-maser emission from the planetary nebula K3-35. The masers are in a magnetized torus with a radius of about 85 astronomical units and are also found at the surprisingly large distance of about 5,000 astronomical units from the star, in the tips of bipolar lobes of gas. The precessing jets from K3-35 are probably involved in the excitation of the distant masers, although their existence is nevertheless puzzling. We infer that K3-35 is being observed at the very moment of its transformation from a giant star to a planetary nebula.

K3-35 is an emission-line nebula that contains a bipolar precessing jet, which is the origin of strong emission from N^+ , emanating from a bright core^{5,6}. These structures are embedded in an envelope of radius $\sim 5''$ ($\sim 25,000$ astronomical units (AU) at 5 kpc; ref. 7) with a dense torus⁶ of radius $\sim 12,000$ AU, oriented with position angle (PA) $\sim 125^\circ$. K3-35 was first proposed to be a planetary nebula in 1965 on the basis of its $H\alpha$ emission⁸. However, there has been some controversy in the past about the nature of this object^{5,9-11}. This controversy was settled when the spectrum of K3-35 was analysed^{6,12}. The presence of strong He^+ emission at $4,686 \text{ \AA}$ (with an intensity about half that of the $H\beta$ emission) implies the presence of a considerable amount of twice-ionized helium—which cannot be produced by young stars^{13,14} but is easily produced by central stars of planetary nebulae^{15,16}. In addition, the observed intensity ratios of forbidden emission lines to hydrogen emission lines require excitation conditions and abundance enrichments that are typical of planetary nebulae^{6,15,16} but impossible to find in young stellar objects^{13,14}. Thus, the spectrum indicates unambiguously that K3-35 is a planetary nebula and not a young stellar object. Moreover, only a planetary-nebula classification for K3-35 accounts for the properties of its OH-maser emission⁹, infrared emission^{6,16}, spectral energy distribution¹⁰ and CO emission¹¹. The He^+ emission from K3-35 (ref. 6) indicates a photoionized core, while the OH-maser emission^{5,9} indicates that K3-35 is indeed a very young planetary nebula¹⁷.

The identification of K3-35 as a planetary nebula is important in light of the detection of water-maser emission towards this object⁹. However, the relatively low angular resolution of those observations ($\sim 40''$) precludes unambiguous establishment of the association

between the water-maser and K3-35. Here we report simultaneous observations of the water-vapour maser and continuum emission at 22 GHz from K3-35. In addition, we also observed continuum emission at 3.6 cm wavelength. These observations were obtained with the Very Large Array (VLA) in its more extended configuration. The high spatial resolution ($0.1''$) of our observations and high accuracy achieved (of the order of milliarcseconds) between the relative positions of the maser and the ionized envelope of K3-35 at 22 GHz allow us to establish unambiguously the existence of water in a planetary nebula (Fig. 1).

The four active water-vapour maser regions detected towards K3-35 are located at significant positions in the object (Fig. 1). Masers in region C are observed towards the bright core, with individual spots located at a projected distance of ~ 85 AU from the intensity peak of the continuum emission. This distance coincides with the formation radius of water masers (~ 10 – 100 AU) in giant envelopes¹⁸. The water-vapour maser spots in region C are located on both sides of the core, along the minor axis of the nebula (Fig. 1), and their orientation is compatible with the major axis of the torus (PA $\approx 125^\circ$; ref. 6). Therefore, masers in region C probably trace the innermost parts of the K3-35 torus. Surprisingly, water-vapour maser emission is also found in regions N, S1 and S2, located at the tips of the bipolar lobes, as traced by the 3.6-cm continuum emission, at $\sim 5,000$ AU from the centre. Water masers are not expected at such an enormous distance from an evolved star, where the physical conditions required to pump the water maser¹⁹ are not expected to exist. Furthermore, K3-35 is, to our knowledge, the first cosmic source in which water-vapour masers have been detected at the tips of bipolar lobes. We note that the active maser regions appear projected on ionized regions, suggesting that the material is only partially ionized, with water molecules surviving in dense clumps that still remain neutral. Water-vapour masers have been detected in some pre-planetary nebulae²⁰⁻²², but the envelope is not ionized in such nebulae.

The timescale over which water molecules can survive destruction by the ionization front is extremely short, ≤ 100 years (ref. 4). Therefore, K3-35 is among the youngest planetary nebulae ever identified, and its central star is the most evolved star ever found in which water molecules still survive. The projected distance to the centre of the water-vapour maser spots in region C sets an upper limit of ~ 85 AU to the radius of the fully photoionized region in which He^+ arises. Assuming an expansion velocity of $\sim 25 \text{ km s}^{-1}$ for this region, as deduced from the He^+ emission line⁶, the dynamical age of the ionized region is ≤ 15 years, implying that K3-35 entered its planetary nebula phase after 1984. We note that He^+ emission at $4,686 \text{ \AA}$ was not present (or was extremely faint) in 1986¹⁶, whereas its intensity in 1998 was comparable to that of $H\beta$ ⁶. However, K3-35 was classified as a planetary nebula 20 years earlier⁸. This apparent contradiction can be explained if, at the time of its discovery, K3-35 was not yet in the planetary nebula phase, but in its pre-planetary nebula phase. The $H\alpha$ emission detected in 1965 was observed with very poor spectral resolution⁸, and most probably was a combination of the $H\alpha$ and strong N^+ emission lines arising from shock excitation from the bipolar jets. Thus, the emission detected in 1965 does not necessarily imply that photoionization was then present. The jets can be traced up to $\sim 17,000$ AU from the centre⁶ and for a typical jet velocity of $\sim 100 \text{ km s}^{-1}$, their dynamical age is ~ 800 years, much older than the photoionized region, implying that collimated ejection and, hence, shock-excited emission was already present in the pre-planetary nebula phase of K3-35.

We have also observed the OH maser lines at 1,665 and 1,667 MHz, which are detected towards the bright core (Fig. 1), confirming the association with the nebula. The 1,665-MHz maser spots are located in a band of radius ~ 630 AU oriented at PA $\approx 125^\circ$, indicating that they arise in the torus. The 1,667-MHz maser spots appear distributed in two main groups, each displaced ~ 800 AU from the centre, oriented along a direction similar to that of the jets.

Strong levels of circular polarization (up to $\sim 50\% \pm 10\%$) are present in the 1,665-MHz line (Fig. 2), while marginal circular polarization ($\sim 12\%$) is observed in the 1,667-MHz line. Strong polarization of the OH masers at these frequencies is due to the Zeeman effect, caused by a magnetic field of the order of milligauss (refs 23, 24). Therefore, the 1,665-MHz data indicate the presence of a magnetic field in K3-35. This result favours magnetic collimation models of outflows in planetary nebulae²⁵. The presence of a magnetic field has also been inferred in the planetary nebula OH0.9+1.3¹⁷ and in the pre-planetary nebula IRAS17150²⁶. The novel result in K3-35 is that the magnetic field is located in the torus, as required by some theoretical models²⁷.

Regions N and S in K3-35 appear related to the jets, not only because of their spatial association, but also because of the kinematics. Masers in these two regions trace a bipolar outflow with the same orientation as the jets, region N being blueshifted with respect to region S. Maser activity in these regions could be caused

by shocks driven by the bipolar jet into the neutral envelope of K3-35. In fact, theoretical models²⁸ show that shocks driven by stellar winds are capable of creating the physical conditions necessary to pump the water maser. The active maser regions are resolved in well organized filaments with clear velocity gradients (Fig. 1), as also found in OH masers in giant stars and pre-planetary nebulae²⁹. Regions N and S1 are filaments oriented nearly perpendicular to the jets. In fact, the post-shock region should be perpendicular to the wind direction³⁰. Region S2, however, is elongated along the southern jet, and presents a complex loop-like structure which could be related to jet precession. Although the jets seem related to the excitation of the distant water-vapour masers, it is not clear why this emission is observed only in regions N and S if the precessing jets are now pointing towards another direction ($PA \approx 65^\circ$, Fig. 1). Moreover, ionizing stellar radiation probably reaches regions N and S, as is the case in other distant parts of the lobes (Fig. 1). Thus, the presence and, perhaps, persistence of water molecules in regions N

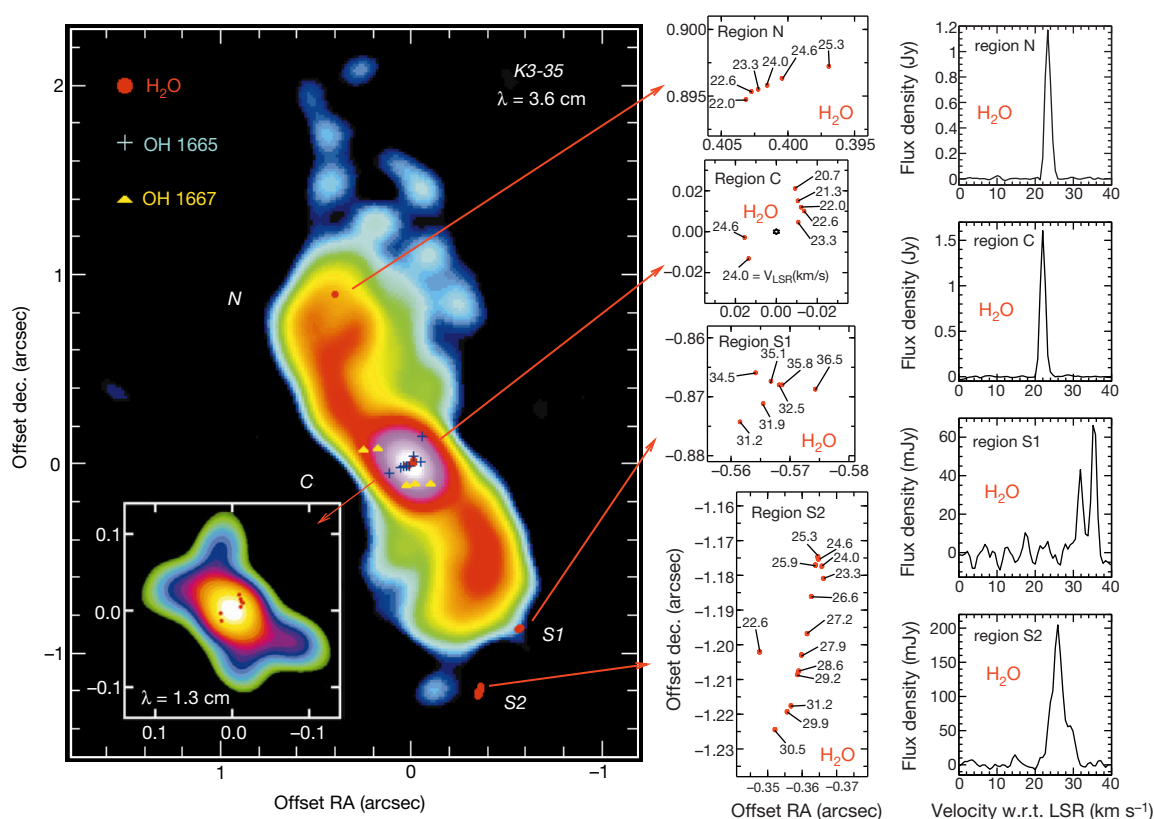


Figure 1 Water-maser and OH-maser emission in the young planetary nebula K3-35. The colour image shows the 3.6-cm continuum emission, which traces the ionized nebular material. A bright core and two bipolar lobes, whose brightest parts correspond to the precessing bipolar jet, can be distinguished. The inset (lower left) shows the 1.3-cm continuum emission corresponding to the core of the nebula. Indications of a bipolar hour-glass-like structure are observed in the 1.3-cm image. Red dots in the figure represent the positions of the H_2O maser components. The detailed spatio-kinematic structure and the spectra of the water-vapour masers are shown in the right-hand panels for the four active maser regions (N, C, S1 and S2) where emission is detected. In the panel of region C, the star marks the peak position of the 1.3-cm continuum emission, at right ascension (J2000) = 19 h 27 min 44.023 s and declination (J2000) = $21^\circ 30' 3.44''$, which is the reference (0, 0) in all the images. The blue crosses mark the positions of the OH 1,665-MHz maser components (velocities with respect to the local standard of rest, LSR, range from 13.4 to 22.2 km s^{-1}), and the yellow triangles mark the positions of the OH 1,667-MHz maser components (LSR velocity is approximately -3 km s^{-1} for the northern group of components, and $\sim 9 \text{ km s}^{-1}$ for the southern group). We note that the OH-maser emission is blueshifted with respect to the water-maser emission. This suggests that OH amplifies the stellar radiation and/or that the redshifted

OH-maser emission may be absorbed by the ionized core. All the continuum and line observations were performed with the VLA in the A configuration on 6 September 1999. Both left and right circular polarizations were observed. At 1.3 cm we observed simultaneously a broad bandwidth of 25 MHz centred at 22.285080 GHz to obtain the continuum data, and a narrower bandwidth of 3.1 MHz with 63 channels of 48.8 kHz each (velocity resolution after Hanning smoothing of $\sim 1.2 \text{ km s}^{-1}$) to observe the H_2O maser line (rest frequency, 22.235080 GHz). The H_2O line data were self-calibrated using the strongest component, and the solutions were applied to both the narrow and broad bandwidth data to cross-calibrate in phase and amplitude. The positional accuracy between the H_2O masers and the 1.3 cm continuum is $\sim 0.001''$. Cleaned maps of the continuum emission at 1.3 cm and 3.6 cm shown in this figure were obtained after reconstruction with circular beams of $0.05''$ and $0.2''$ respectively. Total continuum flux density is $60 \pm 1 \text{ mJy}$ at 1.3 cm, and $38.9 \pm 0.2 \text{ mJy}$ at 3.6 cm. The OH maser lines at 1,665.4018 MHz and 1,667.3590 MHz were observed using a bandwidth of 0.781 MHz and a velocity resolution after Hanning smoothing of $\sim 2.2 \text{ km s}^{-1}$. The angular resolution of the OH observations is $\sim 1''$ and the relative positional accuracy between the OH maser components is $\sim 0.01''$ for the 1,665-MHz data and $\sim 0.05''$ for the 1,667-MHz data.

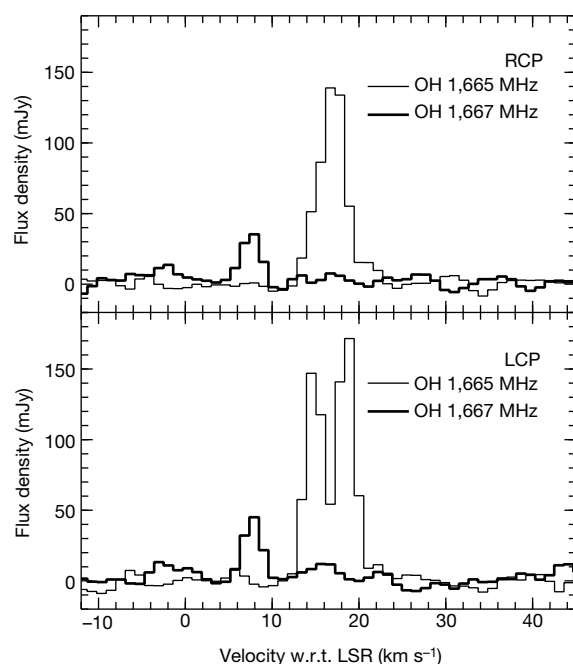


Figure 2 Spectra of the OH 1,665-MHz and 1,667-MHz lines in K3-35. The total velocity range covered by the OH observations was $\sim 140 \text{ km s}^{-1}$. The figure shows the spectra for the left circular polarization (LCP; bottom) and right circular polarization (RCP; top), corresponding to the integrated emission over a region of $< 1''$ in size where all the OH-maser emission detected arises (Fig. 1).

and S requires a shielding mechanism against ionizing radiation. A full understanding of this mechanism will require detailed modelling studies. □

Received 7 June; accepted 27 September 2001.

- Lewis, B. M. The chronological sequence of circumstellar masers: identifying proto-planetary nebulae. *Astrophys. J.* **338**, 234–243 (1989).
- Elitzur, M. Astronomical masers. *Annu. Rev. Astron. Astrophys.* **30**, 75–112 (1992).
- Habing, H. J. Circumstellar envelopes and asymptotic giant branch stars. *Astron. Astrophys. Rev.* **7**, 97–207 (1996).
- Gómez, Y., Moran, J. M. & Rodríguez, L. F. H₂O and SiO maser emission in OH/IR stars. *Rev. Mex. Astron. Astrophys.* **20**, 55–66 (1990).
- Aaquist, O. B. Detailed radio morphology of the compact nebula K3-35. *Astron. Astrophys.* **267**, 260–264 (1993).
- Miranda, L. F. *et al.* High-resolution spectroscopy and broad-band imaging of the young planetary nebula K3-35. *Mon. Not. R. Astron. Soc.* **311**, 748–754 (2000).
- Zhang, C. Y. A statistical distance scale for galactic planetary nebulae. *Astrophys. J. Suppl. Ser.* **98**, 659–678 (1995).
- Kohoutek, L. Hamburg Schmidt-camera survey of faint planetary nebulae. *Bull. Astron. Inst. Czech.* **16**, 221–226 (1965).
- Engels, D., Schmid-Burgk, J., Walmsley, C. M. & Winnberg, A. K3-35: planetary nebula or compact HII region? *Astron. Astrophys.* **148**, 344–346 (1985).
- Aaquist, O. B. & Kwok, S. Bipolar radio morphology in the compact nebula K3-35. *Astron. Astrophys.* **222**, 227–230 (1989).
- Dayal, A. & Bieging, J. H. Millimeter-wave observations of CO in planetary nebulae. *Astrophys. J.* **472**, 703–710 (1996).
- Acker, A. Interactive winds in low-mass evolved stars. *Astrophys. Space Sci.* **260**, 185–198 (1998).
- Evans, I. N. & Dopita, M. A. Theoretical models for HII regions. I. Diagnostic diagrams. *Astrophys. J. Suppl. Ser.* **58**, 125–142 (1985).
- Raga, A. C., Böhm, K.-H. & Cantó, J. A compilation of optical spectrophotometry of HH objects and its tentative interpretation. *Rev. Mex. Astron. Astrofis.* **32**, 161–174 (1996).
- Pottasch, S. R. *Planetary Nebulae* (Reidel, Dordrecht, 1984).
- Acker, A., Marcout, J., Ochsenbein, F., Stenholm, B. & Tyndra, R. *Strasbourg-ESO Catalogue of Galactic Planetary Nebulae* (ESO, Garching, 1992).
- Zijlstra, A. *et al.* OH maser emission from young planetary nebulae. *Astron. Astrophys.* **217**, 157–178 (1989).
- Spencer, J. H., Johnston, K. J., Moran, J. M., Reid, M. J. & Walker, R. C. The structure of H₂O masers associated with late-type stars. *Astrophys. J.* **230**, 449–455 (1979).
- Marvel, K. B. The circumstellar environment of evolved stars as revealed by studies of circumstellar water masers. *Publ. Astron. Soc. Pacif.* **109**, 1286–1287 (1997).
- Bowers, P. F. & Morris, M. The three-dimensional structure of a circumstellar maser. *Astrophys. J.* **276**, 646–652 (1984).

- Likkel, L. & Morris, M. The circumstellar water fountains of IRAS 16342-3814: a very high velocity bipolar outflow. *Astrophys. J.* **329**, 914–919 (1988).
- Marvel, K. B. & Boboltz, D. A. Observations of water masers associated with the proto-planetary nebula candidate IRAS 19296+2227. *Astron. J.* **118**, 1791–1797 (1999).
- Goldreich, P., Keeley, D. A. & Kwan, J. Y. Astrophysical masers II. Polarization properties. *Astrophys. J.* **179**, 111–134 (1973).
- Cohen, R. J. Compact maser sources. *Rep. Prog. Phys.* **52**, 881–943 (1989).
- Blackman, E. G., Frank, A., Markiel, J. M., Thomas, J. H. & Van Horn, H. M. Dynamos in asymptotic-giant-branch stars as the origin of magnetic fields shaping planetary nebulae. *Nature* **409**, 485–487 (2001).
- Hu, J. Y., Slijkhuis, S., Rieu, N.-Q. & de Jong, T. IRAS 17150-3224: a young, optically bipolar, proto-planetary nebula. *Astron. Astrophys.* **273**, 185–193 (1993).
- Rozycka, M. & Franco, J. Toroidal magnetic fields and the evolution of wind-driven nebulae. *Astrophys. J.* **469**, L127–L130 (1996).
- Elitzur, M., Hollenbach, D. J. & McKee, C. F. H₂O masers in star-forming regions. *Astrophys. J.* **346**, 983–990 (1989).
- Zijlstra, A. *et al.* Bipolar outflows in OH/IR stars. *Mon. Not. R. Astron. Soc.* **322**, 280–308 (2001).
- Elitzur, M., Hollenbach, D. J. & McKee, C. F. Planar H₂O masers in star-forming regions. *Astrophys. J.* **394**, 221–227 (1992).

Acknowledgements

We thank L. F. Rodríguez and P. Ho for comments. NRAO is a facility of the National Science Foundation operated under cooperative agreement by Associated Universities, Inc. G.A., L.F.M. and J.M.T. are supported partially by MCYT, Spain. Y.G. acknowledges support from DGAPA, UNAM and CONACyT, Mexico. G.A. acknowledges support from MEC, Spain.

Correspondence and requests for materials should be addressed to L.F.M. (e-mail: lfm@iaa.es).

How many-particle interactions develop after ultrafast excitation of an electron–hole plasma

R. Huber*, F. Tauser*, A. Brodschelm*, M. Bichler†, G. Abstreiter† & A. Leitenstorfer*

* Physik-Department E11, Technische Universität München, James-Frank-Strasse, D-85748 Garching, Germany

† Walter-Schottky-Institut, Technische Universität München, Am Coulombwall, D-85748 Garching, Germany

Electrostatic coupling between particles is important in many microscopic phenomena found in nature. The interaction between two isolated point charges is described by the bare Coulomb potential, but in many-body systems this interaction is modified as a result of the collective response of the screening cloud surrounding each charge carrier^{1,2}. One such system involves ultrafast interactions between quasi-free electrons in semiconductors^{3,4}—which are central to high-speed and future quantum electronic devices. The femtosecond kinetics of non-equilibrium Coulomb systems has been calculated using static^{5,6} and dynamical^{7,8} screening models that assume the instantaneous formation of interparticle correlations. However, some quantum kinetic theories^{9–14} suggest that a regime of unscreened bare Coulomb collisions might exist on ultrashort timescales. Here we monitor directly the temporal evolution of the charge–charge interactions after ultrafast excitation of an electron–hole plasma in GaAs. We show that the onset of collective behaviour such as Coulomb screening and plasmon scattering exhibits a distinct time delay of the order of the inverse plasma frequency, that is, several 10^{-14} seconds.

The plasma state is often called the fourth state of matter, in addition to solids, liquids and gases. In a plasma, charged particles such as electrons and ions or nuclei are no longer bound together as atoms, molecules or solid-state crystals. The charge carriers are free to move and respond independently to electromagnetic fields. In

momentum space, the interaction between two isolated elementary charges e is given by the bare Coulomb potential V_q (ref. 1):

$$V_q = \frac{4\pi e^2}{q^2} \quad (1)$$

We note that V_q depends solely on the momentum exchange q of the two particles during a collision. In a many-body system, the bare Coulomb potential becomes modified as a result of collective effects: each negative charge attracts a cloud of positive charges and vice versa. The combination of a charge carrier and the surrounding screening cloud is commonly referred to as a 'dressed' quasi-particle. The effective screened interaction potential $W_q(\omega, t_D)$ may be expressed as^{1,2}:

$$W_q(\omega, t_D) = \frac{V_q}{\epsilon_q(\omega, t_D)} \quad (2)$$

where V_q is renormalized via the longitudinal dielectric function $\epsilon_q(\omega, t_D)$. $W_q(\omega, t_D)$ depends on the energy $\hbar\omega$ exchanged in a collision. Correspondingly, the dressed interaction is no longer instantaneous. It becomes time-dependent owing to the retarded response of the polarization cloud. The retardation is a result of a resonance of the entire many-body system at the plasma frequency ω_{pl} , which is directly proportional to the square root of the particle pair density N divided by the reduced effective mass m^* . For a nonequilibrium situation, the interaction will also depend upon the time delay t_D with respect to a short perturbation.

Here we address the following question: how quickly are dressed quasi-particles formed, in the case of an initially uncorrelated many-body system of bare particles? Such ultrafast transient phenomena have been suggested to be important in a variety of nonequilibrium systems. Examples range from unscreened positive charges in metals which may influence surface chemistry¹⁵ to the retarded formation of correlations in strongly perturbed nuclear matter¹⁶. We directly observe the build-up of Coulomb correlations and the formation of dressed quasi-particles in a photoexcited electron-hole plasma in the semiconductor GaAs. This model situation has been studied in theory via quantum kinetic calculations for the nonequilibrium many-body system⁹⁻¹⁴. Experimental access to nonequilibrium dynamics in semiconductors is provided by ultrafast laser spectroscopy^{3,4}. Recent experiments exploring the

regime of Coulomb quantum kinetics¹⁷⁻²¹ have probed the particle distributions and interband polarizations. The time evolution of these quantities contains the scattering in an integral manner. It is less sensitive to the exact form of the interaction potential. In order to study the build-up of screening itself, the low-energy excitations of the plasma and the resulting resonances in $\epsilon_q(\omega, t_D)$ must be measured directly with femtosecond temporal resolution ($1 \text{ fs} = 10^{-15} \text{ s}$). For the stationary case of doped semiconductors, these plasmons may be observed with classical infrared²² and light scattering (ref. 23 and see ref. 24 for an overview) techniques. The transient infrared properties of a photoinduced electron-hole plasma have been studied for frequencies below 3 THz ($1 \text{ THz} = 10^{12} \text{ Hz}$)^{25,26} and above 75 THz (ref. 27), with a limited time resolution of 150 fs.

We investigated the evolution of the complex dielectric function renormalizing the bare Coulomb potential (see equation (2)) in the extremely early stage after creation of a nonequilibrium plasma. The principle of our experiment is sketched in Fig. 1. First, free electron-hole pairs are created in a GaAs film via interband transitions

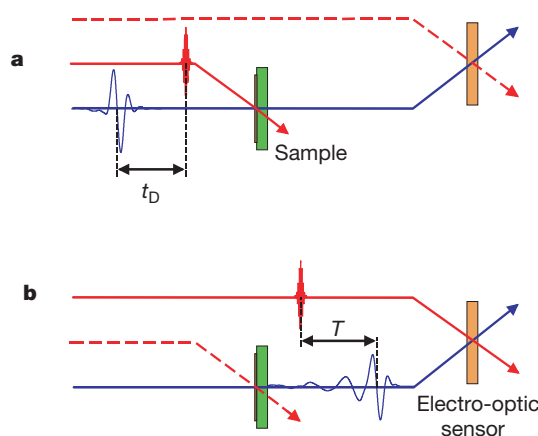


Figure 1 The experimental principle. **a**, First, a thin GaAs sample is excited with a 10-fs laser pulse (red line) at 800 nm, resonant to the interband transition. The sample consists of a 200-nm epitaxial layer of high-purity GaAs (100) on diamond (green block). A single-cycle terahertz electric-field transient (blue line, duration of 27 fs) probes the polarization response of the excited electron-hole plasma after a delay time t_D . The terahertz probe transient is generated via phase-matched optical rectification of a 10-fs visible pulse in a GaSe nonlinear crystal as thin as $30 \mu\text{m}$ (ref. 28). **b**, Second, the transmitted THz field is measured as a function of time T via ultrabroadband electro-optic sampling in a (110) oriented ZnTe crystal (orange block) of a thickness of $10 \mu\text{m}$ (refs 29 and 30).

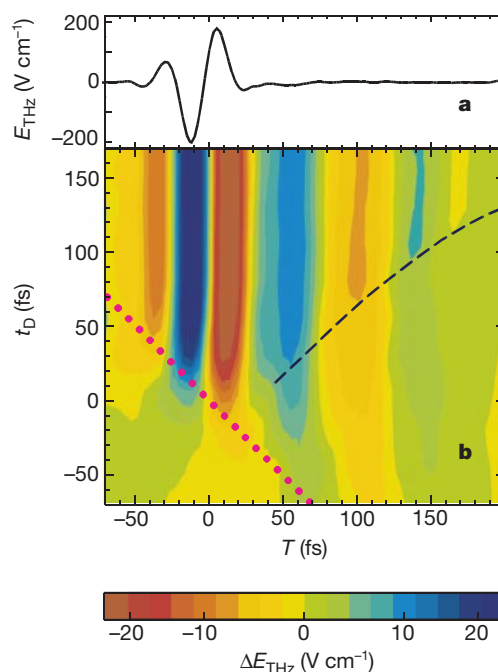


Figure 2 Two-time dependent polarization response of the electron-hole plasma. **a**, Electric-field amplitude E_{THz} of the single-cycle probe versus time T . The transient has been corrected for the frequency response of the electro-optic detector³⁰. The Fourier amplitude spectrum contains components between 1 THz and 100 THz, spanning the entire mid- and far-infrared wavelength region from $\lambda = 300 \mu\text{m}$ to $3 \mu\text{m}$. The phase spectrum is flat between 10 THz and 55 THz. For this frequency interval, the single-cycle pulse represents the ultimate probe with a time resolution limited only by the uncertainty principle. **b**, Electric-field change ΔE_{THz} induced by 10-fs photoexcitation of $2 \times 10^{18} \text{ cm}^{-3}$ electron-hole pairs in GaAs versus delay times t_D and T . The diagonal pink dotted line denotes the position of the maximum of the 10-fs pump pulse. For pump-probe delays between $t_D = -20 \text{ fs}$ and $+20 \text{ fs}$, the excitation pulse overlaps with the electric field of the terahertz probe and a negative (red) and positive (blue) half cycle in ΔE_{THz} appear along the excitation diagonal. These strong features in the vertical region around $T = 0 \text{ fs}$ correspond to an instantaneous perturbation of the single-cycle probe. A retarded oscillatory response of the plasma starts to develop at $t_D = 30 \text{ fs}$, visualized by the dashed black line to guide the eye: An additional half cycle in ΔE_{THz} builds up, represented by a vertical blue column at $T = 60 \text{ fs}$. In this region no probe field is present without excitation (see Fig. 2a). For even longer pump-probe delays $t_D > 70 \text{ fs}$, a second half wave (orange) appears in the retarded response at $T = 100 \text{ fs}$. Finally, another maximum shows up at $T = 140 \text{ fs}$, above the $t_D = 100 \text{ fs}$ horizontal and a last minimum appears at $T = 170 \text{ fs}$ for $t_D \geq 120 \text{ fs}$.

induced with a resonant light pulse (shown red in Fig. 1) of a duration of 10 fs and a central photon energy of 1.55 eV. The photoinduced particle pair density is $N = 2 \times 10^{18} \text{ cm}^{-3}$, resulting in a plasma frequency of $\omega_{\text{pl}}/2\pi \approx 15 \text{ THz}$. The properties of this nonequilibrium system are subsequently probed by a different pulse, a single-cycle electric field transient with a duration of 27 fs (full-width at half-maximum, FWHM, of the intensity) and a centre frequency of 28 THz (sketched blue in Fig. 1)²⁸. This pulse serves as an extremely short probe pulse for the field response of the semiconductor at a time delay t_D after excitation of the plasma and in the limit of small wave vectors q . Any resonance in the system will manifest itself as a retarded 'tail' owing to induced polarization and a phase distortion of the probe transient after transmission through the sample. As we shall see, the changes induced in the probe electric field contain direct information on the many-body effects renormalizing the bare Coulomb interaction. These changes are directly measured in the time domain in a second stage of the experiment (see Fig. 1b). A third laser pulse with a duration of 10 fs analyses the terahertz wave form of the transmitted probe field as a function of a second delay time T via free-space electro-optic sampling^{29,30}.

The observed electric field amplitude E_{THz} of the terahertz probe transmitted through the unexcited GaAs sample versus delay time T is shown in Fig. 2a. The terahertz electric field change ΔE_{THz} due to carrier excitation with the 10-fs pump is displayed in a colour map versus pump-probe delay t_D (vertical) and terahertz sampling delay T (horizontal) in Fig. 2b. The polarization response of the carrier plasma follows instantaneously upon the single-cycle probe pulse at early delay times t_D between -20 fs and 20 fs . With increasing values of t_D the plasma develops a delayed oscillatory response after the terahertz probe burst: starting with a first retarded half wave for $t_D = 30 \text{ fs}$, up to two delayed full oscillation cycles are observed for $t_D \geq 120 \text{ fs}$. At pump-probe delays beyond $t_D = 150 \text{ fs}$, the electric field change ΔE_{THz} versus T becomes stationary on a picosecond timescale determined by the carrier recombination time in the sample.

In order to analyse further the amplitude- and phase-resolved data of Fig. 2 and to extract the complex $\epsilon_{q=0}(\omega, t_D)$, we performed an incomplete Fourier transform^{11,25} into the frequency domain only along the terahertz response axis T . The imaginary and real parts of the deduced inverse dielectric function are shown versus frequency for different pump-probe delays t_D in Fig. 3a and b,

respectively. These data demonstrate the build-up of the dressed interaction in the many-body system with respect to the initial bare Coulomb potential V_q . On top of Fig. 3, the frequency scale has been multiplied by Planck's constant in order to convert into the energy $\hbar\omega$ exchanged in a collision between particles. We start our discussion with the negative imaginary part of $1/\epsilon_{q=0}$, which is a measure for the energy loss of a charged particle propagating in the many-body system. We note that the bare Coulomb interaction of equation (1) lacks an imaginary part. As a consequence, $-\text{Im}(1/\epsilon_{q=0})$ vanishes exactly for the unexcited system (green line in Fig. 3a), except for a small peak at $\omega_{\text{LO}}/2\pi = 8.8 \text{ THz}$, the frequency of the longitudinal optical phonon. This feature comes from the fact that even a single electron in GaAs may exchange the energy quantum $\hbar\omega_{\text{LO}} = 36 \text{ meV}$ with the crystal lattice by polar-optical scattering³¹. When the electron-hole plasma is injected at $t_D = 0 \text{ fs}$, the resonance shifts to higher energies and becomes strongly broadened. The maximum spectral width is obtained at a pump-probe delay of $t_D = 25 \text{ fs}$. At this point, a wide range of energies may be exchanged between the charge carriers, indicating an uncorrelated state in an extreme nonequilibrium situation. Within 100 fs, the broad feature narrows significantly and a sharp resonance in $-\text{Im}(1/\epsilon_{q=0})$ appears at a plasma frequency of $\omega_{\text{pl}}/2\pi = 14.4 \text{ THz}$ (Fig. 3a). We observe the transition to a collective response of the many-body system: The carrier dynamics becomes dominated by scattering with plasmons, the energy quanta of the plasma oscillation, of an energy of $\hbar\omega_{\text{pl}} = 60 \text{ meV}$.

The build-up of a dressed Coulomb interaction is well described by the real part of the inverse dielectric function in Fig. 3b. This quantity renormalizes the charge that an electron is affected by while interacting with another quasi-particle in the plasma, exchanging an energy $\hbar\omega$; see equations (1) and (2). For late pump-probe delays of $t_D > 150 \text{ fs}$, a negative minimum is found for frequencies below the plasma frequency. Obviously, the effective interaction in equation (2) changes sign for an energy exchange slightly below $\hbar\omega_{\text{pl}}$. This phenomenon is due to the oppositely charged screening cloud surrounding the bare particle which is excited resonantly and dominates the collision. In contrast, $\text{Re}(1/\epsilon_{q=0})$ has a positive maximum at energies slightly above $\hbar\omega_{\text{pl}}$: if the plasma is driven above resonance, the polarization oscillates out of phase with the electromagnetic perturbation. As a result, the Coulomb forces due to the bare charge and the screening cloud add up with the same sign. These features are consistent with a fully

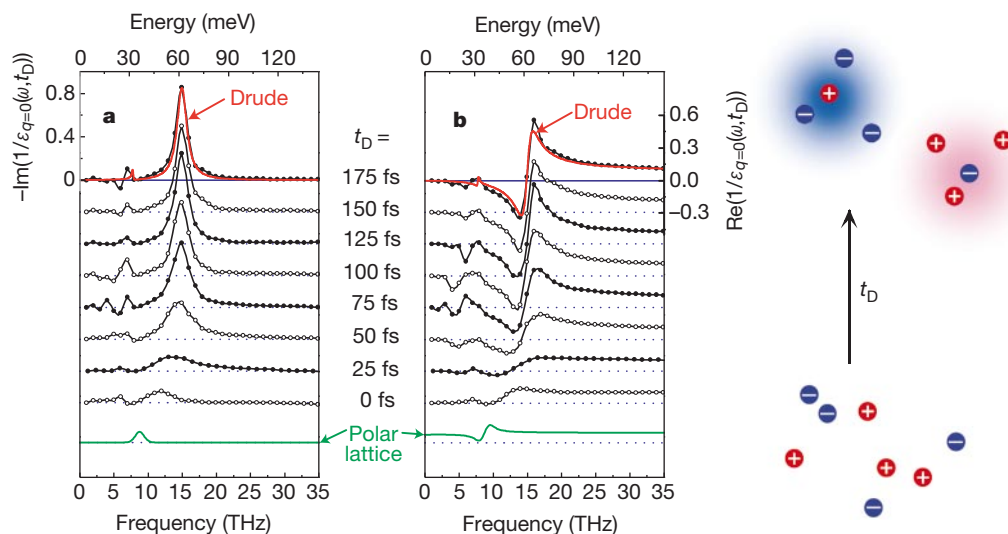


Figure 3 Build-up of Coulomb screening and plasmon scattering. **a, b**, Imaginary (**a**) and real (**b**) part of the long-wavelength limit of the inverse dielectric function of GaAs versus frequency for different pump-probe delays t_D , as deduced from the time-domain data in Fig. 2. The green lines result from a dielectric oscillator model for unexcited GaAs. A

Drude fit of the data at $t_D = 175 \text{ fs}$ ($\tau = 84 \text{ fs}$ and $\omega_{\text{pl}} = 14.4 \text{ THz}$) is drawn red. The sketch on the right visualizes the transition from a chaotic system of bare charges to an ensemble of correlated screened quasi-particles. Dotted lines are offset zero lines (blue).

developed dressed interaction. We note that the extrema around ω_{pl} in $\text{Re}(1/\epsilon_{q=0})$ originating from anti- and over-screening do not emerge instantaneously with carrier generation but show a delayed rise within the first 120 fs. At $t_{\text{D}} = 25$ fs, the spectrum is completely flat above $\hbar\omega_{\text{pl}}$, indicating bare Coulomb collisions without any screening.

The classical model for the long-wavelength and high-frequency limit of the dielectric function of an electron gas is given by Drude theory^{1,32}, assuming an exponential plasmon damping with a time constant τ . We performed least-square fits to our data allowing ω_{pl} and τ as free parameters and keeping the lattice part fixed. Good agreement with a Drude shape is found only for late delay times with $\omega_{\text{pl}} = 14.4$ THz at $t_{\text{D}} = 175$ fs (see red curves in Fig. 3). τ is a measure for the memory depth of the system and for the duration of a collision. It increases from $\tau \approx 20$ fs at $t_{\text{D}} = 25$ fs to $\tau = 85$ fs for $t_{\text{D}} \geq 120$ fs. Our experimental findings support quantum kinetic theories for the nonequilibrium dynamics of the Coulomb interaction^{10–14}. These models predict a delayed build-up of screening and a strongly broadened plasmon pole after ultrafast excitation. The approximate timescale for these phenomena is linked to the duration of a plasma oscillation period which is $2\pi/\omega_{\text{pl}} = 70$ fs in our experiment.

Thus, we have demonstrated a direct test for a basic concept in nonequilibrium many-body physics: the formation of dressed quasi-particles. We show that Coulomb screening and plasmon scattering are not present instantaneously after 10-fs photoexcitation of a dense electron–hole plasma in GaAs, but emerge on a timescale comparable to the inverse plasma frequency. This class of phenomena marks the earliest stage in the dynamics of strongly perturbed particle ensembles far from equilibrium. The results are obtained by applying a new technique, where the polarization response of the system to a single-cycle electric field transient covering the entire mid infrared is directly resolved. The dynamics of elementary excitations in this important spectral region can now be accessed with uncertainty-limited femtosecond resolution of both amplitude and phase. New perspectives arise for investigations in systems such as magnons in high- T_{c} superconductors, lattice dynamics in organic semiconductors, and vibrational relaxation in large molecules and biological complexes. □

Received 14 June; accepted 18 September 2001.

- Mahan, G. D. *Many-Particle Physics* 2nd edn (Plenum, New York, 1993).
- Lindhard, J. On the properties of a gas of charged particles. *Dan. Mat.-fys. Medd.* **28**, 2–57 (1954).
- Shah, J. *Ultrafast Spectroscopy of Semiconductors and Semiconductor Nanostructures* 2nd edn (Springer, Berlin, 1999).
- Chemla, D. S. & Shah, J. Many-body and correlation effects in semiconductors. *Nature* **411**, 549–557 (2001).
- Osman, M. A. & Ferry, D. K. Monte Carlo investigation of the electron–hole-interaction effects on the ultrafast relaxation of hot photoexcited carriers in GaAs. *Phys. Rev. B* **36**, 6018–6032 (1987).
- Goodnick, S. M. & Lugli, P. Effect of electron–electron scattering on nonequilibrium transport in quantum-well systems. *Phys. Rev. B* **37**, 2578–2588 (1988).
- Young, J. F., Henry, N. L. & Kelly, P. J. Full dynamical screening calculation of hot electron scattering rates in multicomponent semiconductor plasma. *Solid State Electron.* **32**, 1567–1571 (1989).
- Collet, J. H. Dynamical screening in the cooling theory of high-density electron–hole plasmas. *Phys. Rev. B* **39**, 7659–7665 (1989).
- Haug, H. & Jauho, A.-P. *Quantum Kinetics in Transport and Optics of Semiconductors* (Springer, Berlin, 1996).
- Hartmann, M., Stolz, H. & Zimmermann, R. Kinetics of screening in optically excited semiconductors. *Phys. Status Solidi B* **159**, 35–42 (1990).
- El Sayed, K., Schuster, S., Haug, H., Herzel, F. & Henneberger, K. Subpicosecond plasmon response: Buildup of screening. *Phys. Rev. B* **49**, 7337–7344 (1994).
- Bányai, L., Vu, Q. T., Mieck, B. & Haug, H. Ultrafast quantum kinetics of time-dependent RPA-screened coulomb scattering. *Phys. Rev. Lett.* **81**, 882–885 (1998).
- Kwong, N.-H. & Bonitz, M. Real-time Kadanoff–Baym approach to plasma oscillations in a correlated electron gas. *Phys. Rev. Lett.* **84**, 1768–1771 (2000).
- Vu, Q. T. & Haug, H. Time-dependent screening of the carrier–phonon and carrier–carrier interactions in non-equilibrium systems. *Phys. Rev. B* **62**, 7179–7185 (2000).
- Schöne, W.-D. & Ekardt, W. Time-dependent screening of a positive charge distribution in metals: Excitons on an ultrashort time scale. *Phys. Rev. B* **62**, 13464–13471 (2000).
- Köhler, H. S. Memory and correlation effects in nuclear collisions. *Phys. Rev. C* **51**, 3232–3239 (1995).
- Camescasse, F. X. *et al.* Ultrafast electron redistribution through Coulomb scattering in undoped GaAs: Experiment and theory. *Phys. Rev. Lett.* **77**, 5429–5432 (1996).
- Hügel, W. A. *et al.* Photon echoes from semiconductor band-to-band continuum transitions in the regime of Coulomb quantum kinetics. *Phys. Rev. Lett.* **83**, 3313–3316 (1999).

- Bolton, S. R., Neukirch, U., Sham, L. J., Chemla, D. S. & Axt, V. M. Demonstration of sixth-order Coulomb correlations in a semiconductor single quantum well. *Phys. Rev. Lett.* **85**, 2002–2005 (2000).
- Vu, Q. T., Haug, H., Hügel, W. A., Chatterjee, S. & Wegener, M. Signature of electron–plasmon quantum kinetics in GaAs. *Phys. Rev. Lett.* **85**, 3508–3511 (2000).
- Bonitz, M., Lampin, J. F., Camescasse, F. X. & Alexandrou, A. Nonequilibrium plasmons in optically excited semiconductors. *Phys. Rev. B* **62**, 15724–15734 (2000).
- Spitzer, W. G. & Fan, H. Y. Determination of optical constants and carrier effective masses of semiconductors. *Phys. Rev.* **106**, 882–890 (1957).
- Mooradian, A. & Wright, G. B. Observation of the interaction of plasmons with longitudinal optical phonons in GaAs. *Phys. Rev. Lett.* **16**, 999–1001 (1966).
- Abstreiter, G., Cardona, M. & Pinczuk, A. Light scattering by free carrier excitations in semiconductors. *Springer Topics Appl. Phys.* **54**, 5–150 (1984).
- Schall, M. & Jepsen, P. U. Photoexcited GaAs surfaces studied by transient terahertz time-domain spectroscopy. *Opt. Lett.* **25**, 13–15 (2000).
- Beard, M. C., Turner, G. M. & Schmuttenmaer, C. A. Transient photoconductivity in GaAs as measured by time-resolved terahertz spectroscopy. *Phys. Rev. B* **62**, 15764–15777 (2000).
- Ganikhanov, F., Burr, K. C., Hilton, D. J. & Tang, C. L. Femtosecond optical-pulse-induced absorption and refractive-index changes in GaAs in the midinfrared. *Phys. Rev. B* **60**, 8890–8896 (1999).
- Huber, R., Brodschelm, A., Tauser, F. & Leitenstorfer, A. Generation and field-resolved detection of femtosecond electromagnetic pulses tunable up to 41 THz. *Appl. Phys. Lett.* **76**, 3191–3193 (2000).
- Wu, Q. & Zhang, X.-C. Free-space electro-optic sampling of mid-infrared pulses. *Appl. Phys. Lett.* **71**, 1285–1286 (1997).
- Leitenstorfer, A., Hunsche, S., Shah, J., Nuss, M. C. & Knox, W. H. Detectors and sources for ultrabroadband electro-optic sampling: Experiment and theory. *Appl. Phys. Lett.* **74**, 1516–1518 (1999).
- Fürst, C., Leitenstorfer, A., Laubereau, A. & Zimmermann, R. Quantum kinetic electron–phonon interaction in GaAs: Energy non-conserving scattering events and memory effects. *Phys. Rev. Lett.* **78**, 3733–3736 (1997).
- Yu, P. Y. & Cardona, M. *Fundamentals of Semiconductors* (Springer, Berlin, 1996).

Acknowledgements

We thank H. Haug, W. Kaiser, L. V. Keldysh and A. Laubereau for discussions and continuous support.

Correspondence and requests for materials should be addressed to A.L. (e-mail: aleitens@ph.tum.de).

On-chip natural assembly of silicon photonic bandgap crystals

Yurii A. Vlasov^{*†}, Xiang-Zheng Bo[‡], James C. Sturm[‡] & David J. Norris^{*}

^{*}NEC Research Institute, 4 Independence Way, Princeton, New Jersey 08540, USA

[†]A. F. Ioffe Physical-Technical Institute, St. Petersburg, Russia

[‡]Department of Electrical Engineering and Center for Photonics and Optoelectronic Materials, Princeton University, Princeton, New Jersey 08544, USA

Photonic bandgap crystals can reflect light for any direction of propagation in specific wavelength ranges^{1–3}. This property, which can be used to confine, manipulate and guide photons, should allow the creation of all-optical integrated circuits. To achieve this goal, conventional semiconductor nanofabrication techniques have been adapted to make photonic crystals^{4–9}. A potentially simpler and cheaper approach for creating three-dimensional periodic structures is the natural assembly of colloidal microspheres^{10–15}. However, this approach yields irregular, polycrystalline photonic crystals that are difficult to incorporate into a device. More importantly, it leads to many structural defects that can destroy the photonic bandgap^{16,17}. Here we show that by assembling a thin layer of colloidal spheres on a silicon substrate, we can obtain planar, single-crystalline silicon photonic crystals that have defect densities sufficiently low that the bandgap survives. As expected from theory, we observe unity reflectance in two crystalline directions of our photonic crystals around a wavelength of 1.3 micrometres. We also show that additional fabrication steps, intentional doping and patterning, can be performed, so demonstrating the potential for specific device applications.

The strategy behind the colloidal assembly approach is to exploit the tendency of monodisperse submicrometre spheres to sponta-

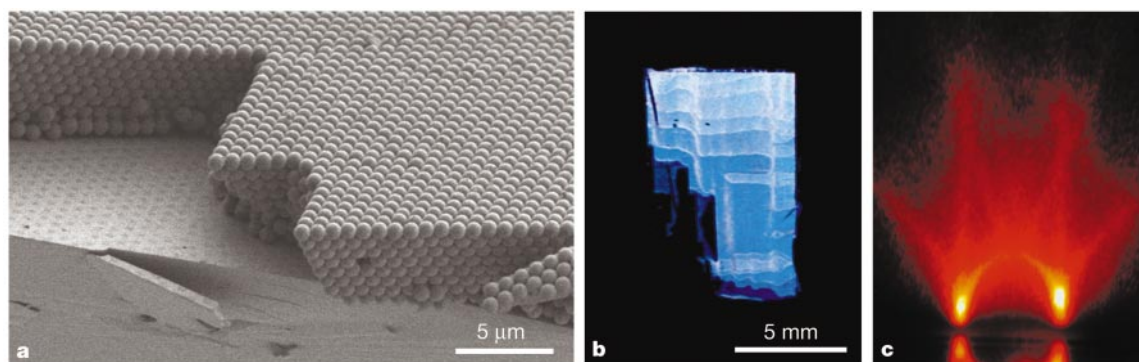


Figure 1 Characterization of thin planar opal templates assembled directly on a Si wafer from 855-nm spheres. **a**, Cross-sectional SEM image. **b**, Large-scale optical photograph, looking down on the wafer. The opal is formed as the meniscus is swept from right to left. The horizontal lines represent monolayer steps in the crystal (lighter shades of blue represent single-layer increases in thickness). **c**, Optical diffraction pattern obtained

from the sample in **b**. Under normal incidence, a characteristic 6-spot diffraction pattern with weak Kossel rings (enhanced in the photo) remained unchanged while scanning a 100- μm -diameter laser spot over centimetre-scale regions of the opal. These 6 spots arise from diffraction off {110} planes of a single-crystal f.c.c. lattice with 1% stacking faults.

neously organize on a face-centred cubic (f.c.c.) lattice. The resulting material—a synthetic opal—acts as a template into which a semiconductor material is infiltrated. Subsequent removal of the template leads to a three-dimensional (3D) photonic crystal—inverted opal—that has periodic air-spheres embedded inside the semiconductor. If the refractive index of the semiconductor is sufficiently high (>2.85), such a structure has been predicted to exhibit a bandgap^{18,19}. However the standard method for growing the initial opal (sedimentation of spheres from suspension) yields centimetre-scale pieces of polycrystalline material with numerous defects in the crystal lattice (stacking faults, dislocations and point defects). As the bandgap in f.c.c. photonic crystals is relatively narrow, these defects can easily close the gap by filling it with

localized photonic states¹⁶.

We used an alternative method to form synthetic opals. Recent research has improved control over colloidal crystallization in various geometries^{20–22}. In particular, strong capillary forces at a meniscus between a substrate and a colloidal sol can induce crystallization of spheres into a 3D array of controllable thickness²². If this meniscus is slowly swept across a vertically placed substrate by solvent evaporation, thin planar opals can be deposited. As solvent evaporation must compete with sedimentation, this method is believed to be limited to spheres with diameters $<0.4\ \mu\text{m}$. But spheres with larger diameters ($\sim 0.8\ \mu\text{m}$) are required to make photonic crystals with a bandgap at technologically important wavelengths such as 1.3 or 1.5 μm (ref. 19), so we added a convective

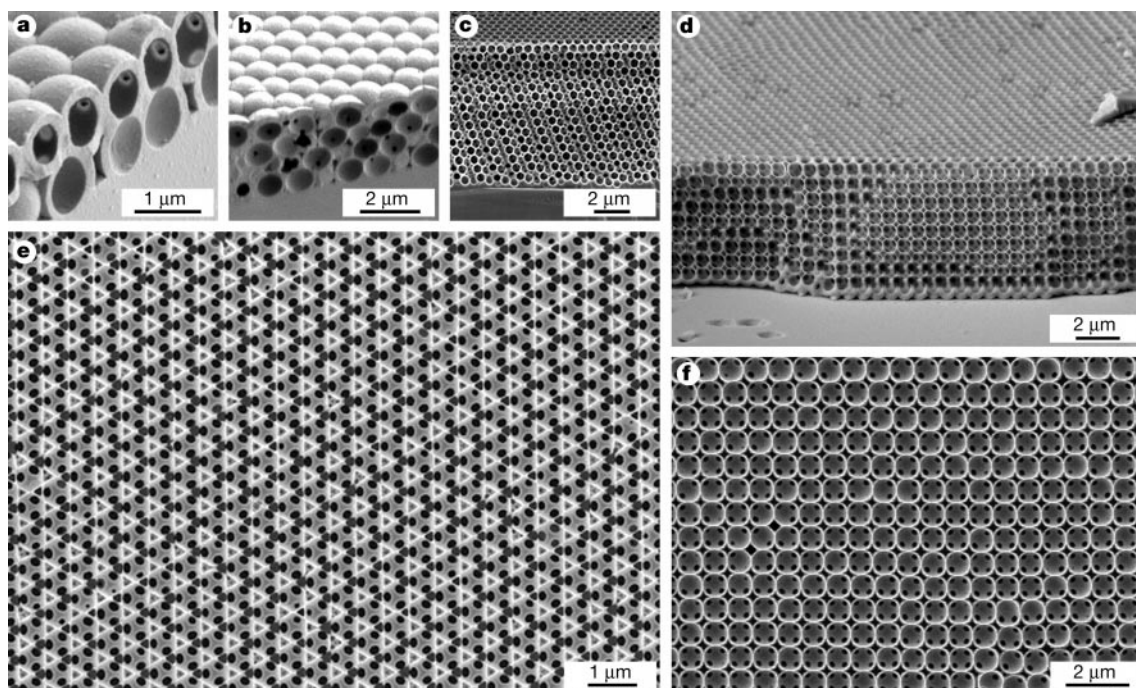


Figure 2 SEM images of planar Si photonic crystals. Cross-sectional SEM images are shown as a function of the thickness of the initial opal template for 2 (**a**), 4 (**b**), and 16 (**c**) layers. The Si substrate is snapped, and the fracture propagates up through the photonic crystal. Close examination of the lowest layer reveals that the photonic crystals are completely integrated into the wafer. Apparent defects in **c** are due to a ripple in the fracture surface that occurs in thicker photonic crystals. **d**, Sample edge showing the

(100) surface, confirming that the crystals are f.c.c. **e**, A planar (111) surface exposed by RIE. From this image, parameters¹⁹, such as the radius of the Si coating sphere (0.41) and the radius of the air sphere (0.354), both in units of the lattice constant, were estimated. **f**, A planar (100) surface exposed by RIE. The sphere diameters are 670 nm (**c** and **e**), 855 nm (**d** and **f**), and 1 μm (**a** and **b**).

flow to the sol to minimize sedimentation and provide a continuous flow of particles toward the meniscus region. We found straightforward conditions that yielded planar opals using large (up to $1\ \mu\text{m}$) silica colloids. For example, a Si wafer was placed vertically in a vial containing an ethanolic suspension of silica spheres ($855 \pm 1.3\%$ in size, 1% by volume). Flow was achieved by placing a temperature gradient across the vial (from 80°C at the bottom to 65°C near the top). Scanning electron microscopy (SEM) images of the resulting templates (Fig. 1a) indicate that the defect densities ($\sim 1\%$ stacking faults, $\sim 10^{-3}$ point defects per unit cell) are much lower than for sedimented opals ($\sim 20\%$ stacking faults, $\sim 10^{-2}$ point defects per

unit cell¹⁶). In addition, this approach yields large-sphere opals up to 20 layers thick that coat centimetre-scale areas of the Si wafer (Fig. 1b). SEM and optical diffraction measurements (Fig. 1c) show that these opals have single crystalline domains ($1\ \text{mm}$ – $1\ \text{cm}$) that are 10–100 times larger than in the best sedimented opals. We speculate that the improved quality of these samples is due to a meniscus-induced shear that aligns the close-packed layers into a f.c.c. crystal during deposition, as shown in other geometries²³.

Once such a template was prepared, its interstitial spaces were filled with Si to satisfy the refractive index requirement for the photonic bandgap. In previous work, a purpose-built apparatus was required, in which disilane was first condensed into the pores of the opal at cryogenic temperatures and subsequently decomposed by heating at pressures of 200 torr (ref. 15). Although homogeneous infiltration was demonstrated for sedimented opals, this process did not allow sufficient control over the deposition to fill our thin planar opals. Instead, we filled planar opals using a commercially available low-pressure chemical vapour deposition (LPCVD) furnace that provides complete control of the growth parameters²⁴. As LPCVD is surface-reaction-limited, this technique is in principle well suited to conformal filling of the interstitials of the opal. Furthermore, an advantage of LPCVD is that it is the standard Si deposition technique for the microelectronics industry (used, for example, in complementary metal-oxide-semiconductor, CMOS, technology). Unfortunately, under typical CMOS fabrication conditions near 600°C , filling the opal template can be problematic. First, infiltration of the structure can be limited by premature obstruction of the outermost channels ($\sim 100\ \text{nm}$) of the opal, which provide gas transport to the innermost layers. Second, deposition results in polycrystalline silicon (poly-Si) with grains ($\sim 100\ \text{nm}$) that can introduce undesirable roughness at surfaces inside the final photonic crystal. By decreasing the temperature to 550°C , we obtained homogeneous infiltration with LPCVD even for templates as thick as 40 layers. The lower temperature reduced the sticking coefficient of the precursor, allowing deposition to penetrate all the way to the Si wafer without a visible interface (Fig. 2a). Temperatures below 580°C also avoided internal surface roughness by uniformly depositing amorphous silicon (a-Si), which was then transformed into a poly-Si structure with smooth interfaces by annealing at 600°C for 8 hours. After deposition, the silica template was removed by wet etching. Thus thin planar inverted opals of controllable thickness were obtained (Fig. 2a–d) that were incorporated directly into the wafer and inherited the advantageous mechanical properties of poly-Si.

We explored the existence of the photonic bandgap in such crystals using optical spectroscopy. As structural defects can eliminate the bandgap in Si inverted opals and most defects are transferred from the template, we first examined the photonic properties of the unfilled opal (Fig. 3a and b). Figure 3b shows the photonic band diagram calculated for unfilled opals along the [111] direction of the f.c.c. lattice. The optical response of the samples was measured in the region of the higher order bands, where the diagram becomes quite complex owing to numerous interband interactions. Such experiments are important because the influence of disorder increases at higher energies where the optical wavelength becomes comparable to even small imperfections in the lattice. Further, for structures with higher refractive index the bandgap eventually develops in this region. Figure 3a shows the transmission spectrum for a 7-layer-thick opal template prepared on a transparent quartz substrate. This result, plotted on an absolute scale, was obtained with a microscope-based spectrometer attached to a two-dimensional (2D) CCD (charge-coupled device) array²⁵. While the apparatus allows transmission spectra to be obtained with a spatial resolution of $\sim 2\ \mu\text{m}$, such spectra were quite reproducible over the entire field of view of the CCD ($\sim 90\ \mu\text{m}$). When compared with calculations based on the multiple-scattering technique²⁶ (Fig. 3a), the experimental results reproduce the position and intensity of

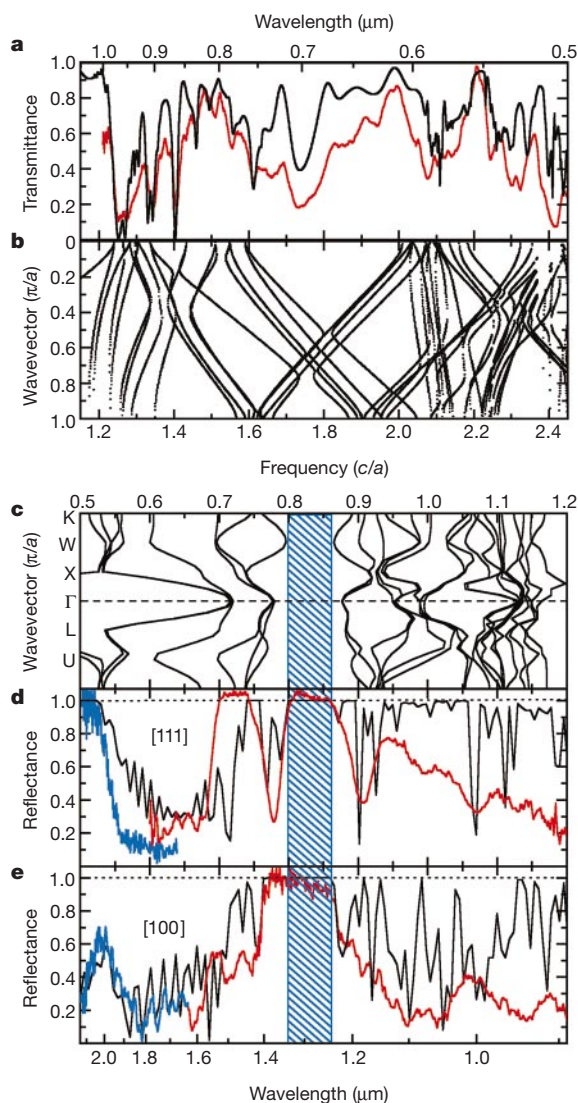


Figure 3 Comparison of optical results with calculations. **a**, Experimental (red) and calculated (black) transmission spectra for incidence normal to the (111) plane of a 7-layer planar opal made from 855-nm silica spheres with a refractive index of 1.45. The frequency is plotted in units of c/a , where c is the speed of light and a is the lattice constant. **b**, The photonic band diagram calculated along the [111] direction for the same parameters used in **a**. **c**, The photonic band diagram calculated for the Si inverted opal measured in **d** and **e**. **d**, Experimental reflection spectra for amorphous Si inverted opals measured normal to the (111) plane. Data from two samples, with a equal to 1,070 nm (red) and 841 nm (blue), are combined. The wavelength scale corresponds to the 1,070-nm sample. The best theoretical fit (black) was obtained for a Si coating sphere radius of 0.428 and an air sphere radius of 0.354. **e**, Experimental and theoretical reflection spectra as in **d** for incidence normal to the (100) plane. The blue-hatched region denotes the expected frequency range of the bandgap. For calculations, we used 12.25 as the dielectric constant of Si.

all predicted spectral features up to the fiftieth photonic band (frequency ~ 2.4). This strongly supports the SEM evidence that the defect density in our opal templates is markedly reduced.

To perform analogous optical measurements on our Si photonic crystals, reactive ion etching (RIE) was first used to eliminate the Si coating ($0.2\text{--}0.4\text{ }\mu\text{m}$) on the top surface of the samples, which can introduce unwanted photonic surface states. By performing selective RIE before removal of the template, the silica spheres automatically terminated etching one half-layer into the crystal. Thus, a planar (111) surface, initially buried inside the structure, was exposed (Fig. 2e). Examination of the fine details in this image allowed the structural parameters of the crystal to be estimated (see Fig. 2). Using these parameters, we calculated the photonic band structure as shown in Fig. 3c²⁷. The bandgap, which in inverted opals occurs between the eighth and ninth bands (the blue-hatched region in Fig. 3), was centred around $1.3\text{ }\mu\text{m}$ in our samples. To optically probe this region, we replaced the CCD in our microscope set-up with an InGaAs photodiode. Spectra were obtained (Fig. 3d and e) either by scanning the monochromator or by replacing it with a Fourier transform infrared spectrometer. Absolute reflectance was determined with an experimental accuracy of $\pm 5\%$ by comparing the sample reflection with that from a silver mirror. Typical reflectance spectra, measured normal to the (111) surface for 8-layer-thick samples, are presented in Fig. 3d. Near-unity reflectance is observed both in the region of the lower-order stop band (frequency ~ 0.5) and in the vicinity of the bandgap (frequency ~ 0.83) in good agreement with our calculations using the Translight computer code based on the transfer matrix method²⁸. At higher frequencies the experimental reflectance decreases owing to the onset of absorption in Si.

The crucial test of the bandgap is to examine whether this high reflectance is maintained for directions other than [111]. Fortunately, with some effort, crystalline regions of [100] orientation can be found in our samples (Fig. 2f). Although these regions, which form near monolayer steps in the crystal²⁹, are very small, our microscope set-up allows their optical response to be obtained. In Fig. 3e reflection spectra are presented for such [100]-oriented regions, 8 layers thick. As in the [111] data (Fig. 3d), the bandgap region is characterized by a broad, 'square-top' reflection peak with near-unity reflectance. Furthermore, comparison of Fig. 3d and e shows that a broad range of frequencies exists around $1.3\text{ }\mu\text{m}$ where

near-unity reflectivity is maintained for both directions, as expected by theory for the bandgap. Thus, with our approach we can obtain planar Si bandgap crystals grown directly on a Si wafer.

To demonstrate the potential usefulness of these crystals, we performed two additional steps: doping and patterning. Intentional doping of the lattice with artificial defects is now possible because the concentration of native defects in our crystals is significantly reduced. Artificial defects are necessary to construct low-threshold lasers and light-emitting diodes as they can act as microcavities of high quality factor (Q). In comparison to 2D structures where the Q of such cavities is inherently limited by the out-of-plane leakage of light, 3D bandgap crystals can completely confine the light to obtain the highest Q ($>10^4$) while maintaining a diffraction-limited modal volume. In Fig. 4a, we demonstrate the first step toward creating such microcavities in our crystals. A trace amount of silica spheres of another size was added to the initial opal template before Si deposition. Consequently, an additional air cavity was introduced into one of the unit cells of the lattice. By changing the size of the dopant spheres, this cavity can be tuned to trap photons in photonic acceptor states within the bandgap. For real devices, the placement of such defects—which in Fig. 4a are randomly distributed throughout the structure—would need to be controlled. Methods to achieve this by more sophisticated assembly of the initial opal template are currently being explored.

Another step is to pattern the crystal for a specific device requirement. The planar configuration of our samples allows the possibility of using standard semiconductor microfabrication techniques, such as photolithography and RIE. As natural assembly has already defined the complex nanoscale topology of the 3D crystal, photolithography is more than sufficient to pattern the crystal into a microscale device. To demonstrate the feasibility of this idea, Fig. 4b shows an array of $100\text{-}\mu\text{m}$ -diameter 'rings' of photonic crystal that are three layers thick; the array covers a 1-mm region of the wafer. This pattern was first formed in a spun-on layer of photoresist by photolithography, and then transferred into the photonic crystal by RIE. Thicker samples can be processed by anisotropic RIE with high aspect ratios, as shown in Fig. 4c and d. Densely packed arrays of bandgap crystal devices could, in principle, be formed on a single chip. Furthermore, the technological steps required to fabricate our structures (LPCVD, oxide removal by wet etching, photolithography and RIE) are standard Si-based microfabrication techniques.

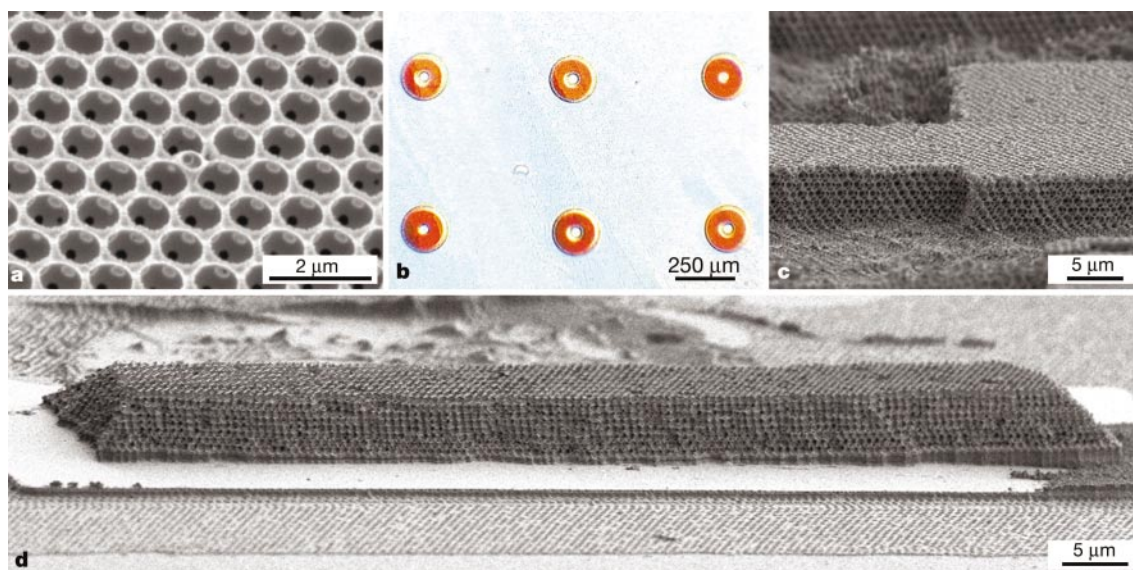


Figure 4 Doping and patterning Si photonic crystals. **a**, SEM image showing the introduction of an interstitial defect. **b**, Photograph of an array of $100\text{-}\mu\text{m}$ photonic-crystal rings patterned into a large-area crystal via photolithography and etching. The red

colour arises owing to optical diffraction from the crystal. **c**, **d**, Cross-sectional SEM images of patterned photonic crystals.

So, this approach could allow the eventual mass production of low-cost, naturally assembled, Si bandgap devices incorporated into microelectronic integrated circuits.

Such integration is desirable for adding optical functionality to current Si microelectronic devices. Devices made from planar Si bandgap crystals could allow on-chip manipulation of photons. Furthermore, the possibility of creating light in the same Si devices also exists³⁰. If such an active material could be incorporated inside a high-Q microcavity in our structure, a route to Si micro-lasers could be obtained. Thus, the above materials may provide an inexpensive route to all-silicon integrated optoelectronic circuits, using present-day technology. □

Received 19 June; accepted 9 October 2001.

- Joannopoulos, J. D., Villeneuve, P. R. & Fan, S. Photonic crystals: putting a new twist on light. *Nature* **386**, 143–149 (1997).
- Yablonovitch, E. Inhibited spontaneous emission in solid-state physics and electronics. *Phys. Rev. Lett.* **58**, 2059–2062 (1987).
- John, S. Strong localization of photons in certain disordered dielectric superlattices. *Phys. Rev. Lett.* **58**, 2486–2489 (1987).
- Krauss, T. F., De La Rue, R. M. & Brand, S. Two-dimensional photonic-bandgap structures operating at near-infrared wavelengths. *Nature* **383**, 699–702 (1996).
- Chow, E. *et al.* Three-dimensional control of light in a two-dimensional photonic crystal slab. *Nature* **407**, 983–986 (2000).
- Smith, C. J. M. *et al.* Low-loss channel waveguides with two-dimensional photonic crystal boundaries. *Appl. Phys. Lett.* **77**, 2813–2815 (2000).
- Painter, O. *et al.* Two-dimensional photonic band-gap defect mode laser. *Science* **284**, 1819–1821 (1999).
- Lin, S. Y. *et al.* A three-dimensional photonic crystal operating at infrared wavelengths. *Nature* **394**, 251–253 (1998).
- Noda, S., Tomoda, K., Yamamoto, N. & Chutinan, A. Photonic bandgap crystals at near-infrared wavelengths. *Science* **289**, 604–606 (2000).
- Astratov, V. N. *et al.* Optical spectroscopy of opal matrices with CdS embedded in its pores: quantum confinement and photonic band gap effects. *Nuovo Cimento D* **17**, 1349–1354 (1995).
- Wijnhoven, J. E. G. J. & Vos, W. L. Preparation of photonic crystals made of air spheres in titania. *Science* **281**, 802–804 (1998).
- Vlasov, Y. A., Yao, N. & Norris, D. J. Synthesis of photonic crystals for optical wavelengths from semiconductor quantum dots. *Adv. Mater.* **11**, 165–169 (1999).
- Braun, P. V. & Wiltzius, P. Electrochemically grown photonic crystals. *Nature* **402**, 603–604 (1999).
- Müller, M., Zentel, R., Maka, T., Romanov, S. G. & Sotomayor-Torres, C. M. Photonic crystal films with high refractive index contrast. *Adv. Mater.* **12**, 1499–1503 (2000).
- Blanco, A. *et al.* Large-scale synthesis of a silicon photonic crystal with a complete three-dimensional bandgap near 1.5 micrometres. *Nature* **405**, 437–440 (2000).
- Vlasov, Y. A. *et al.* Manifestation of intrinsic defects in optical properties of self-organized opal photonic crystals. *Phys. Rev. E* **61**, 5784–5793 (2000).
- Li, Z.-Y. & Zhang, Z.-Q. Fragility of photonic band gaps in inverse-opal photonic crystals. *Phys. Rev. B* **62**, 1516–1519 (2000).
- Sözüer, H. S., Haus, J. W. & Inguva, R. Photonic bands: convergence problems with the plane-wave method. *Phys. Rev. B* **45**, 13962–13972 (1992).
- Busch, K. & John, S. Photonic band gap formation in certain self-organizing systems. *Phys. Rev. E* **58**, 3896–3908 (1998).
- Denkov, N. D. *et al.* Two-dimensional crystallization. *Nature* **361**, 26 (1993).
- van Blaaderen, A., Ruel, R. & Wiltzius, P. Template-directed colloidal crystallization. *Nature* **385**, 321–324 (1997).
- Jiang, P., Bertone, J. F., Hwang, K. S. & Colvin, V. L. Single-crystal colloidal multilayers of controlled thickness. *Chem. Mater.* **11**, 2132–2140 (1999).
- Amos, R. M., Rarity, J. G., Tapster, P. R., Shepherd, T. J. & Kitson, S. C. Fabrication of large-area face-centered-cubic hard-sphere colloidal crystals by shear alignment. *Phys. Rev. E* **61**, 2929–2935 (2000).
- Kamins, T. *Polycrystalline Silicon for Integrated Circuits and Displays* 2nd edn, 10–22 (Kluwer, Boston, 1998).
- Vlasov, Y. A., Deutch, M. & Norris, D. J. Single domain spectroscopy of self-assembled photonic crystals. *Appl. Phys. Lett.* **76**, 1627–1629 (2000).
- Stefanou, N., Yannopoulos, V. & Modinos, A. Heterostructures of photonic crystals: frequency bands and transmission coefficients. *Comput. Phys. Commun.* **113**, 49–77 (1998).
- Johnson, S. G. & Joannopoulos, J. D. Block-iterative frequency-domain methods for Maxwell's equations in a planewave basis. *Opt. Express* **8**, 173–190 (2001).
- Bell, P. M., Pendry, J. B., Martin-Moreno, L. & Ward, A. J. A program for calculating photonic band structures and transmission coefficients of complex structures. *Comput. Phys. Commun.* **85**, 306–322 (1995).
- Dushkin, C. D., Lazarov, G. S., Kotsev, S. N., Yoshimura, H. & Nagayama, K. Effect of growth conditions on the structure of two-dimensional latex crystals: experiment. *Colloid Polym. Sci.* **277**, 914–930 (1999).
- Ng, W. L. *et al.* An efficient room-temperature silicon-based light-emitting diode. *Nature* **410**, 192–194 (2001).

Acknowledgements

We thank A. Reynolds for providing transfer matrix computer code (Translight), P. Chaikin and S. Wagner for discussions, and N. Yao for experimental assistance. This work was partially supported by DARPA/ONR.

Correspondence and requests for materials should be addressed to D.J.N. (e-mail: dnorris@research.nj.nec.com).

Smart single-chip gas sensor microsystem

C. Hagleitner, A. Hierlemann, D. Lange, A. Kummer, N. Kerness, O. Brand & H. Baltes

Physical Electronics Laboratory, ETH Zurich, Hönggerberg, HPT H 6, CH-8093 Zurich, Switzerland

Research activity in chemical gas sensing is currently directed towards the search for highly selective (bio)chemical layer materials, and to the design of arrays consisting of different partially selective sensors that permit subsequent pattern recognition and multi-component analysis^{1–3}. Simultaneous use of various transduction platforms has been demonstrated^{4–6}, and the rapid development of integrated-circuit technology has facilitated the fabrication of planar chemical sensors^{7,8} and sensors based on three-dimensional microelectromechanical systems^{9,10}. Complementary metal-oxide silicon processes have previously been used to develop gas sensors based on metal oxides¹¹ and acoustic-wave-based sensor devices¹². Here we combine several of these developments to fabricate a smart single-chip chemical microsensor system that incorporates three different transducers (mass-sensitive, capacitive and calorimetric), all of which rely on sensitive polymeric layers to detect airborne volatile organic compounds. Full integration of the microelectronic and micromechanical components on one chip permits control and monitoring of the sensor functions, and enables on-chip signal amplification and conditioning that notably improves the overall sensor performance. The circuitry also includes analog-to-digital converters, and an on-chip interface to transmit the data to off-chip recording units. We expect that our approach will provide a basis for the further development and optimization of gas microsystems.

Physisorption and bulk dissolution of analyte molecules within the polymer volume constitute the predominant interaction mechanisms of polymer-based sensors¹³. The physical properties of the polymer change on absorption of analytes: in the system we report here, these changes are detected by three distinct complementary metal-oxide silicon (CMOS) microtransducers (Fig. 1) in response to fundamentally different molecular properties of the analytes. One transducer (Fig. 1b) responds to the mass of sorbed molecules, another responds to the heat of absorption (Fig. 1c), and the third—the capacitive sensor—responds to a combination of the volume and dielectric properties of the absorbates convoluted with changes in these parameters for the sensing layer (Fig. 1a).

The capacitive sensor relies on interdigitated electrode structures, which correspond to the two plates of a standard capacitor, to monitor changes of the dielectric coefficient of the polymer. One electrode is made of the first metal layer of the CMOS process (E2 in Fig. 1a), whereas the second electrode is a stack of two metal layers (E1 in Fig. 1a). This is to enhance the sensitivity of the capacitive microsystem by increasing the number of electric field lines within the polymer volume by applying a 'three-dimensional' design¹⁴. The dimensions of the capacitor are 800 × 800 μm, and its electrode width and spacing are 1.6 μm. As the nominal capacitance of the interdigitated capacitor is a few picofarads, and the expected capacitance changes on absorption of a volatile organic compound are in the attofarad range, a dedicated on-chip measurement configuration and specific signal-conditioning circuitry had to be developed. The sensor response is read out as a differential signal between a polymer-coated sensing capacitor and a passivated reference capacitor. A digital output signal is then generated by comparing the minute loading currents of both capacitors using fully differential second-order sigma-delta modulator circuitry¹⁴. The final output is a digital word with a precision of 20 bits.

Within the capacitive sensor, the space containing 95% of the electric field includes a polymer volume that is within a distance from the transducer surface of half the periodicity of the electrodes. For a layer thickness of less than half the periodicity, swelling of the polymer on analyte absorption always results in a capacitance increase, regardless of the dielectric constant of the absorbed analyte. This is due to the increased polymer/analyte volume within the field-line region exhibiting a larger dielectric constant than that of the air it displaces. On the other hand, the capacitance change for a polymer layer thickness larger than half the periodicity of the electrodes is determined by the ratio of dielectric constants of analyte and polymer. If the dielectric constant of the polymer is lower than that of the analyte, the capacitance will be increased, and if the polymer dielectric constant is larger, the capacitance will be decreased. This effect has been previously detailed and supported by simulations¹⁵.

The second type of transducer is a micromachined cantilever, which has been developed in the context of atomic force microscopy^{16–19} (AFM). Resonating cantilevers are a very promising type of mass-sensitive chemical sensor^{20–22}, especially if properly co-integrated with circuitry²³. The absorption of analyte in the chemically sensitive polymer causes shifts in resonance frequency as a consequence of changes in the oscillating mass (Fig. 1b). Our CMOS resonant chemical sensor is based on a 150- μm -long cantilever, which consists mainly of silicon, with some vapour-deposited oxide and thermal oxide. The different thermal expansion coefficients of the silicon and the oxide (bimorph effect) are used to initiate a cantilever vibration by periodically applying electric pulses to heating resistors embedded at the cantilever base. The cantilevers resonate at their fundamental mechanical frequency of 380 kHz with a quality factor of approximately 1,000 in air. The cantilever vibration is detected by embedded piezoresistors in a Wheatstone-bridge configuration. An on-chip low-noise fully differential difference amplifier with a gain of 30 dB amplifies the output signal of the Wheatstone bridge. The cantilever acts as the frequency-determining element in a feedback oscillation circuit, which was (for the first time, to our knowledge) entirely integrated on the chip with a counter. For more details on the electronics, see ref. 23.

The third transducer is a thermoelectric calorimeter based on the Seebeck effect, which detects enthalpy changes on absorption (heat of condensation) or desorption (heat of vaporization) of analyte molecules in the polymer film^{24,25} (Fig. 1c). These enthalpy changes cause temperature variations on thermally insulated micro-machined membranes. An array of 256 polysilicon/aluminium thermocouples connected in series is sandwiched in between the dielectric membrane layers to record these temperature variations. Their hot junctions are located on the membrane, the cold junctions

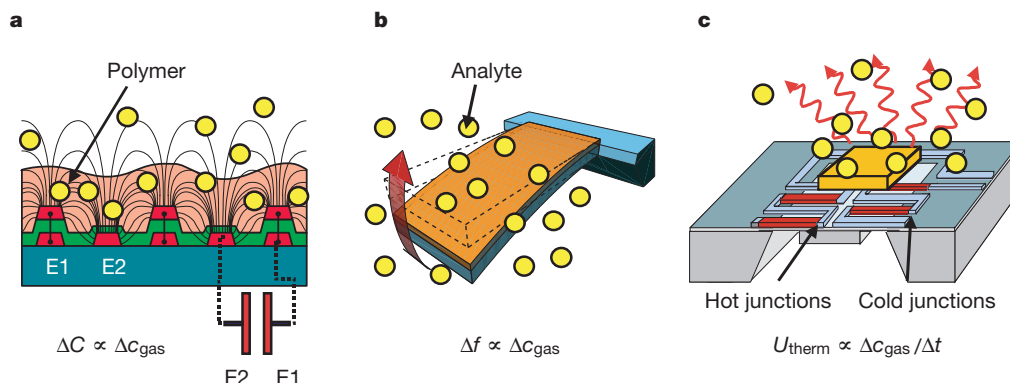


Figure 1 Principles of three different types of transducer. **a**, Microcapacitor sensitive to changes in dielectric properties (capacitance change, ΔC , proportional to analyte concentration change, ΔC_{gas}); **b**, resonant cantilever sensitive to mass changes (frequency change, Δf , proportional to analyte concentration change, ΔC_{gas}); and

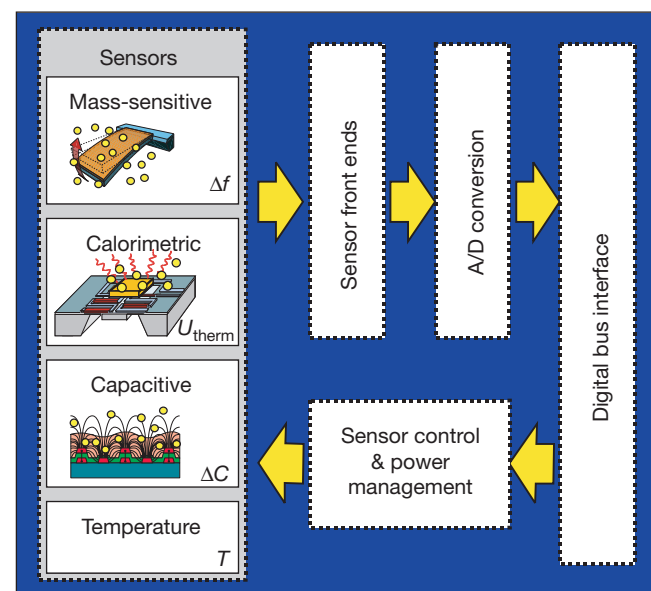


Figure 2 Diagram of the overall microsystem architecture. It consists of sensors, driving and signal-conditioning circuitry (sensor front end), analog-to-digital converters, sensor control and power management unit, and a digital interface.

on the bulk chip. A thermovoltage between the hot and cold junctions develops, which is proportional to the temperature difference (in the millikelvin range). The calorimetric sensing process thus includes four principal steps: (1) absorption of the analyte and subsequent partitioning, (2) liberation or abstraction of heat, which causes (3) temperature changes, which are (4) transformed into thermovoltage changes^{26,27}. In contrast to the capacitive and gravimetric chemical sensors relying on equilibrium signals, the calorimetric sensor detects only transients, meaning only changes in the analyte concentration^{24–27}.

The calorimetric sensor system includes two $500 \times 500 \mu\text{m}$ dielectric membranes, one of which is coated with the gas-sensitive polymer, while the other is left uncoated and serves as a reference²⁸. The sensing and reference thermopiles are connected in parallel to the input stage of a low-noise, chopper-stabilized, instrumentation amplifier on-chip (gain 8,000, bandwidth 500 Hz) to record the temperature differences between the two membranes. The thermovoltage is translated into a digital signal on-chip using a sigma-delta analog-to-digital converter and a decimation filter.

The overall system chip also includes a temperature sensor,

c, microcalorimeter measuring the absorption or desorption heat upon interaction of organic volatiles with the polymer (thermoelectric voltage, U_{therm} , proportional to analyte concentration gradient $\Delta C_{\text{gas}} / \Delta t$).

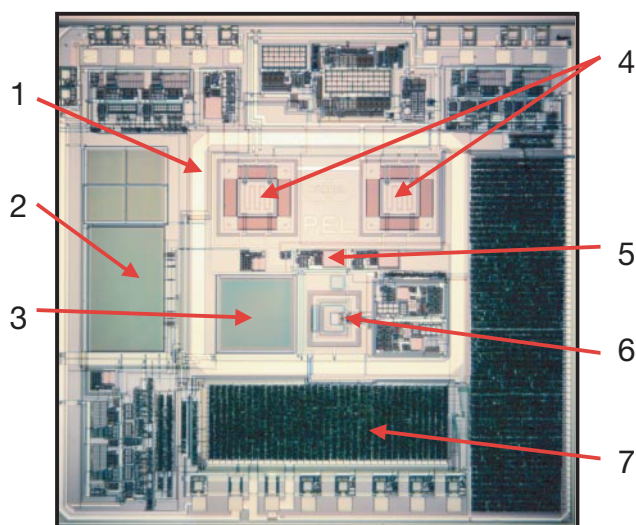


Figure 3 Micrograph of the gas microsensor system chip (size, 7×7 mm). The different components include: 1, flip-chip frame; 2, reference capacitor; 3, sensing capacitor; 4, calorimetric sensor and reference; 5, temperature sensor; 6, mass-sensitive resonant cantilever; and 7, digital interface.

because bulk physisorption of volatiles in polymers is strongly temperature-dependent: a temperature increase of 10°C decreases the fraction of absorbed analyte molecules by approximately 50%, and thus leads to a drastic sensor signal reduction. The operation temperature must be exactly known to enable quantitative measurements. The temperature sensor relies on the linear temperature dependence of a bipolar transistor available in the CMOS process. The voltage is converted to a digital signal using a sigma-delta converter. After calibration, the temperature sensor exhibits an accuracy of 0.1 K at operation temperatures between 223 and 383 K .

A diagram of the overall microsensor system architecture is shown in Fig. 2, and a chip micrograph is displayed in Fig. 3. On the left-hand side of Fig. 2, we have the four different sensors. The sensor front ends represent all the sensor-specific driving circuitry and signal-conditioning circuitry, as already described in the context of the different transducers. The analog-to-digital conversion is done on-chip as well. This leads to a high signal-to-noise ratio, because noisy connections are avoided and a robust digital signal is generated on-chip and then transmitted to an off-chip data port via an I^2C digital interface (I^2C is a communication standard developed by Philips, Eindhoven, The Netherlands). The I^2C bus interface

offers the additional advantage of having only very few signal lines (essentially two) for bi-directional communication, and also allows the operation of multiple chips on the same bus system. An on-chip digital controller manages the sensor timing and the chip power budget.

Figure 3 shows the processed and tested single-chip gas sensor system. The overall chip size is $7 \times 7\text{ mm}$. The chip processing steps include an unaltered 15-mask commercial CMOS process (provided by austriamicrosystems, Unterpremstätten, Austria), followed by the application of a back-side mask and anisotropic potassium hydroxide etching of the membrane structures for the calorimetric sensors and the cantilever. Front-side reactive ion etching is then applied to release the cantilever from the respective membrane. Finally, the cantilever, calorimeter and capacitor sensors are coated with the polymer using an airbrush method. The sensors are located in the centre of a metal frame, which is used to apply a flip-chip packaging technique²⁹.

Figure 4 displays simultaneously recorded sensor signals of the three transducers upon exposure to $1,200$ and $3,000\text{ p.p.m.}$ of ethanol, and $1,000$ and $3,000\text{ p.p.m.}$ of toluene, at 301 K . The sensors were alternately exposed to analyte-loaded gas and pure carrier gas. The polymeric coating consisted of poly(etherurethane)²⁷, at a thickness of approximately $4\text{ }\mu\text{m}$. Figure 4a shows the measured frequency signals (output of the sigma-delta converter) of the capacitor. Ethanol—with a dielectric constant of 24.5 , which is larger than that of poly(etherurethane) (2.9)—causes a capacitance increase and hence positive frequency shifts, and toluene (with a dielectric constant of 2.4) causes a capacitance decrease and hence negative frequency shifts. The limits of detection are $3\text{--}5\text{ p.p.m.}$ for ethanol and $5\text{--}8\text{ p.p.m.}$ for toluene.

Figure 4b displays the cantilever response. Ethanol shows rather low signals as compared to toluene, owing to its lower molecular mass and its lower enrichment (partitioning) in the polymeric phase. The extent of physisorption is, to first order, inversely proportional to the boiling temperature of the analyte (ethanol, 351 K ; toluene, 360 K): the lower the boiling temperature, the less enrichment in the polymer¹³. Here, the limits of detection are $10\text{--}12\text{ p.p.m.}$ for ethanol and $1\text{--}2\text{ p.p.m.}$ for toluene, which is comparable to other mass-sensitive transducers³⁰.

The calorimetric results in Fig. 4c represent a superposition of already discussed partitioning and heat-budget change due to analyte ab/desorption. The absorbing analyte liberates heat (predominantly heat of condensation) causing a positive transient signal (positive peak), whereas the desorbing analyte abstracts vaporization heat from the environment, generating a negative transient signal (negative peak upon purging, see Fig. 4c). The inset in Fig. 4c shows the time-resolved response during the first six seconds of

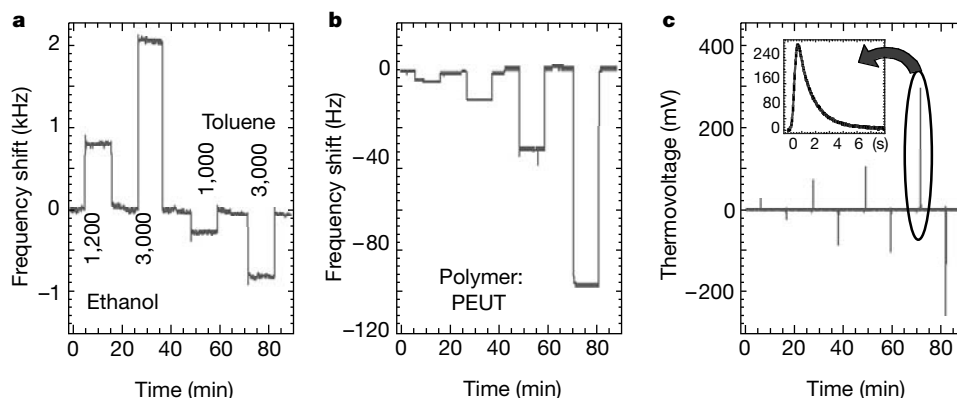


Figure 4 Sensor signals simultaneously recorded from all three polymer-coated (poly(etherurethane), PEUT) transducers. The transducers were exposed to $1,200$ and $3,000\text{ p.p.m.}$ of ethanol, and $1,000$ and $3,000\text{ p.p.m.}$ of toluene, all at 301 K .

a, Frequency shifts (Sigma-Delta converter output) of the capacitor; **b**, frequency shifts of

the resonating cantilever; and **c**, thermovoltage transients of the calorimetric sensor. The inset in **c** is an expanded-scale view, showing the development of the calorimetric transient within 6 s .

exposure to 3,000 p.p.m.. The condensation/vaporization heat of ethanol is larger than that of toluene (42.3 compared to 38.0 kJ mol⁻¹ at 298 K) owing to directional interaction in the liquid phase (compare, for example, water), whereas the polymer partitioning of toluene is stronger owing to its higher boiling point. The limits of detection of the calorimetric method are larger than those of the other transducers (40–50 p.p.m. toluene, 100–150 p.p.m. ethanol).

The signals of all three transducers linearly correlate with analyte concentration at low concentrations (less than 3% of saturation vapour pressure at the operating temperature)¹³. As can be seen from the considerable variation in the transducer responses upon gas exposure, different molecular properties or different aspects of the coating–molecule interaction are measured by the different transducers. Alcohols, for example, provide comparably low signals on mass-sensitive transducers owing to their high saturation vapour pressure and low molecular mass, but provide large signals on capacitors owing to their large dielectric constant (24.5). Drastic changes in thermovoltages on the thermopiles are measured for chlorinated hydrocarbons (not shown here) used, for example, in cooling sprays, which in turn have a low dielectric constant, thus showing only small signals on capacitors.

Another powerful feature of the system is that even a zero response from one of the transducers, if there is a measurable response from the other two, is a highly informative data point to help identify the chemical species. To further improve analyte identification/quantification, an array of microsystem chips coated with different polymers could be used.

Our monolithic CMOS gas microsystem is intended to identify organic solvents in transport containers, or to provide information on workplace safety—in, for example, the chemical industry. It will form part of a hand-held or credit-card-sized detection unit. The CMOS approach offers substantial advantages, such as full microelectronics compatibility, extremely small size, low power consumption, and production at industrial standards. We expect that further research and the fast progress of microelectronics development will improve the device performance in the near future. □

Received 22 May; accepted 10 September 2001.

- Massart, D. L., Vandeginste, B. G. M., Deming, S. N., Michotte, Y. & Kaufman, L. *Chemometrics: a Textbook* Vol. 2 (Data Handling in Science and Technology, Elsevier, Amsterdam, 1988).
- Brereton, R. G. (ed.) *Multivariate Pattern Recognition in Chemometrics* Vol. 9 (Data Handling in Science and Technology, Elsevier, Amsterdam, 1992).
- Hierlemann, A. et al. in *Sensors Update* Vol. 2 (eds Baltes, H., Göpel, W. & Hesse, J.) 119–180 (VCH, Weinheim, 1996).
- Snow, A. W., Barger, W. R., Klusty, M., Wohltjen, H. & Jarvis, N. L. Simultaneous electrical-conductivity and piezoelectric mass measurements on iodine-doped phthalocyanine Langmuir-Blodgett films. *Langmuir* **2**(4), 513–519 (1986).
- Rodriguez, J. L., Hughes, R. C., Corbett, W. T. & McWhorter, P. J. in *Technical Digest IEEE International Electron Devices Meeting* 521–524 (IEEE, New York, 1992).
- Gumbrecht, W. et al. Integrated pO₂, pCO₂, pH sensor system for online blood monitoring. *Sensors Actuators B* **18–19**, 704–708 (1994).
- Van den Berg, A., van der Waal, P. D., van der Schoot, B. B. & de Rooij, N. F. Silicon-based chemical sensors and chemical analysis systems. *Sensors Materials* **6**, 23–43 (1994).
- Müller, G., Deimel, P. P., Hellmich, W. & Wagner, C. Sensor fabrication using thin-film-on-silicon approaches. *Thin Solid Films* **296**, 157–163 (1997).
- Kovacs, G. T. A. *Micromachined Transducers* (WCB McGraw-Hill, Boston, 1998).
- Madou, M. *Fundamentals of Microfabrication* (CRC, Boca Raton, Florida, 1997).
- Suehle, J. S., Cavicchi, R. E., Gaitan, M. & Semancik, S. Tin oxide gas sensor fabricated using CMOS micro-hotplates and *in-situ* processing. *IEEE Electron Device Lett.* **14**, 118–120 (1993).
- Vellekoop, M. J., Lubking, G. W., Sarro, P. M. & Venema, A. Integrated-circuit-compatible design and technology of acoustic-wave-based microsystems. *Sens. Actuators A* **44**, 249–263 (1994).
- Hierlemann, A., Ricco, A. J., Bodenhofer, K., Dominik, A. & Göpel, W. Conferring selectivity to chemical sensors via polymer side-chain selection: Thermodynamics of vapor sorption by a set of polysiloxanes on thickness-shear mode resonators. *Anal. Chem.* **72**, 3696–3708 (2000).
- Koll, A., Kummer, A., Brand, O. & Baltes, H. Discrimination of volatile organic compounds using CMOS capacitive chemical microsystems with thickness-adjusted polymer coating. *Proc. SPIE Smart Struct. Mater.* **3673**, 308–317 (1999).
- Steiner, F. P. et al. in *Digest 8th Int. Conf. on Solid-state Sensors and Actuators* Vol. 2, 814–817 (Foundation for Sensor and Actuator Technology, Stockholm, 1995).
- Gimzewski, J. K., Gerber, C., Meyer, E. & Schlittler, R. R. Observation of a chemical reaction using a micromechanical sensor. *Chem. Phys. Lett.* **217**, 589–594 (1994).
- Stowe, T. D. et al. Attonewton force detection using ultrathin silicon cantilevers. *Appl. Phys. Lett.* **71**, 288–290 (1997).
- Lange, D. et al. in *Proc. IEEE Workshop on Micro Electro Mechanical Systems (MEMS 99)* 447–452 (IEEE, Piscataway, 1999).

- Chen, G. Y., Thundat, T., Wachter, E. A. & Warmack, R. J. Adsorption-induced surface stress and its effects on resonance frequency of microcantilevers. *J. Appl. Phys.* **77**, 3618–3622 (1995).
- Thundat, T., Chen, G. Y., Warmack, R. J., Allison, D. P. & Wachter, E. A. Vapor detection using resonating microcantilevers. *Anal. Chem.* **67**, 519–521 (1995).
- Lang, H. P. et al. A chemical sensor based on a micromechanical cantilever array for the identification of gases. *Appl. Phys. A* **66**, 61–64 (1998).
- Maute, M. et al. Detection of volatile organic compounds with polymer-coated cantilevers. *Sensors Actuators B* **58**, 505–511 (1999).
- Hagleitner, C., Lange, D., Brand, O., Hierlemann, A. & Baltes, H. in *Digest of Technical Papers IEEE International Solid State Circuits Conf. San Francisco* (ed. Wuorinen, J. H.) Vol. 44, 246 (IEEE, Piscataway, 2001).
- Bataillard, P., Steffgen, E., Haemmerli, S., Manz, A. & Widmer, H. M. An integrated silicon thermopile as biosensor for the thermal monitoring of glucose, urea and penicillin. *Biosensors Bioelectron.* **8**, 89–98 (1993).
- Lerchner, J., Seidel, J., Wolf, G. & Weber, E. Calorimetric detection of organic vapors using inclusion reactions with organic coating materials. *Sensors Actuators B* **32**, 71–75 (1996).
- Van Heerwarden, A. W., Sarro, P. M., Gardner, J. W. & Bataillard, P. Liquid and gas micro-calorimeters for (bio)chemical measurements. *Sensors Actuators A* **43**, 24–30 (1994).
- Hierlemann, A. et al. Application-specific sensor systems based on CMOS chemical microsystems. *Sensors Actuators B* **70**, 2–11 (2000).
- Koll, A. et al. in *Proc. IEEE Workshop on Micro Electro Mechanical Systems (MEMS 99)* 547–551 (IEEE, Piscataway, 1999).
- Koll, A. et al. A flip-chip-packaged CMOS chemical microsystem for detection of volatile organic compounds. *Proc. SPIE Smart Struct. Mater.* **3328**, 223–232 (1998).
- Ballantine, D. S. et al. *Acoustic Wave Sensors: Theory, Design, and Physico-chemical Applications* (Academic, San Diego, 1997).

Acknowledgements

We acknowledge the contributions of former staff of the Physical Electronics Laboratory at ETH Zurich who were involved in the development of the chemical microsensor. We also thank the prototype manufacturers austriamicrosystems for their services. This work formed part of a cooperative project involving the University of Tübingen (U. Weimar), the University of Bologna (M. Rudan), and ETH Zürich, which is financially supported by the Körber Foundation, Hamburg, Germany.

Correspondence and requests for materials should be addressed to A.H. (e-mail: hierlema@iqe.phys.ethz.ch).

Formation of coastline features by large-scale instabilities induced by high-angle waves

Andrew Ashton, A. Brad Murray & Olivier Arnault*

Division of Earth and Ocean Sciences/Center for Nonlinear and Complex Systems, Duke University, Box 90227, Durham, North Carolina 27708-0227, USA

Alongshore sediment transport that is driven by waves is generally assumed to smooth a coastline. This assumption is valid for small angles between the wave crest lines and the shore, as has been demonstrated in shoreline models¹. But when the angle between the waves and the shoreline is sufficiently large, small perturbations to a straight shoreline will grow^{2,3}. Here we use a numerical model to investigate the implications of this instability mechanism for large-scale morphology over long timescales. Our simulations show growth of coastline perturbations that interact with each other to produce large-scale features that resemble various kinds of natural landforms, including the capes and cusate forelands observed along the Carolina coast of southeastern North America. Wind and wave data from this area support our hypothesis that such an instability mechanism could be responsible for the formation of shoreline features at spatial scales up to hundreds of kilometres and temporal scales up to millennia.

When waves approach a shore at an oblique angle, nearshore

* Present address: École Normale Supérieure, 45 rue d'ULM, 75005 Paris, France.

wave-breaking produces a shore-parallel current. On a sandy shoreline, this current transports sediment. This alongshore sediment flux, Q_s , is a nonlinear function of the local shoreline angle relative to the wave crests, exhibiting a maximum (Fig. 1a and Methods section). The instability in shoreline shape occurs when waves approach at a relative angle greater than that which maximizes sediment transport (which we will call 'high-angle' waves). In this case, moving alongshore in the transport direction ('downdrift'), Q_s will decrease along the crest of a perturbation as the angle between the shoreline and the waves becomes progressively farther from the transport-maximizing angle (Fig. 1b). This convergence of sediment flux causes accretion; the perturbation will grow. (This conclusion involves the common assumption that the rate of any cross-shore sediment exchange between the nearshore region and deeper water is negligible compared to the alongshore flux.)

Finite-amplitude effects will eventually dictate the evolution of a growing feature. For the case of high-angle waves approaching from a constant direction, as the amplitude of the feature increases, sediment flux at the inflection point downdrift of the crest will approach zero. This will lead to the extension of a spit-like protuberance. In addition, the angle between the shoreline and the wave crests at the inflection point on the updrift side of a feature will approach, but not increase beyond, the angle that maximizes sediment transport. However, the inevitably continued erosion updrift of this point will cause the inflection point to migrate continually towards the crest. This erosion of the updrift flank and accumulation at the crest and downdrift of it will cause the feature to translate downdrift while growing. (If the relative angle between the waves and the regional shoreline trend is only slightly greater than the transport-maximizing angle, subtle, low-amplitude shoreline features will result.)

In the case of multiple shoreline features interacting along an extended shoreline, further morphological evolution resulting from

the above-mentioned instability cannot be predicted analytically. To investigate the long-term effect of this instability, we have developed a numerical model. This model is similar to others that discretize a semi-empirical relationship for alongshore sediment flux versus breaking-wave height and angle relative to the shore (Methods)^{4,5}. Our model, however, is designed to address morphological evolution on large temporal and spatial scales, and can accommodate arbitrary shoreline shapes (Fig. 2, Methods). It is not designed to simulate the details of any particular geographical location, but to investigate more generally how shorelines might respond to high-angle waves.

In the simplest simulations, waves approach an initially nearly straight shoreline from a constant high angle. With this steady, spatially uniform forcing, the simple, deterministic interactions in the model produce increasingly complex shapes (Fig. 3a). But such simulations omit an important natural condition: varying wave angles. In most simulations, wave angles are distributed according to a probability distribution function (Methods). Features grow for all distributions weighted towards high-angle waves. When the distribution is asymmetric, causing a net sediment flux in one alongshore direction, simulated features translate alongshore (Fig. 3b). In this case, features can overtake each other and merge, and their average wavelength and amplitude increase with time.

Somewhat surprisingly, the wavelength and amplitude of the shoreline forms also continually increase for the case of a symmetric wave distribution (Fig. 3c). Simulations reveal that slightly larger features shelter their neighbours from the highest-angle waves, causing the larger features to grow relative to the smaller features, increasing the sheltering effect. The smaller features will eventually experience a wave distribution weighted towards low-angle waves, resulting in the rapid disappearance of the smaller features and growth of their neighbours. As the features develop, the wave

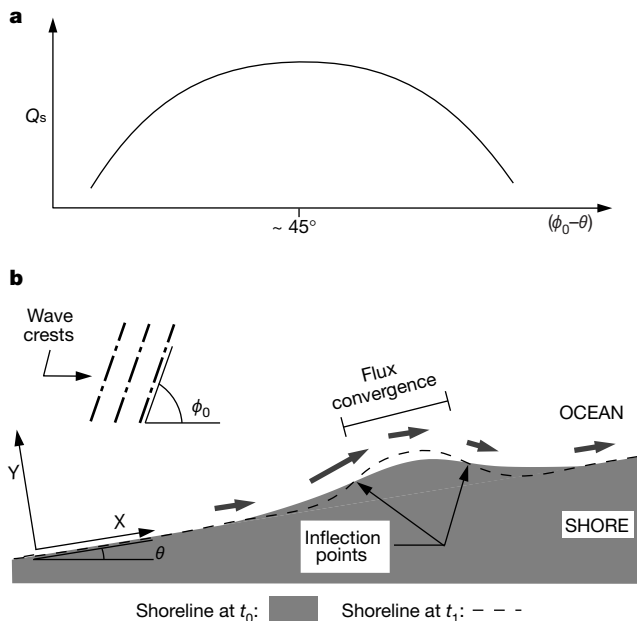


Figure 1 As a result of the basic instability, shoreline perturbations grow in the presence of high-angle waves. **a**, Schematic relationship between alongshore sediment flux, Q_s , and the relative angle between wave crests in deep water (before nearshore refraction), ϕ_0 , and the local shoreline orientation, θ , showing a maximum for $\phi_0 - \theta \approx 45^\circ$. **b**, The relative magnitude of the alongshore sediment flux, shown by the length of the arrows, and the consequent zones of erosion and accretion on a perturbation to a straight shoreline when the angle between the wave crests and the shoreline trend is greater than that which maximizes the sediment flux.

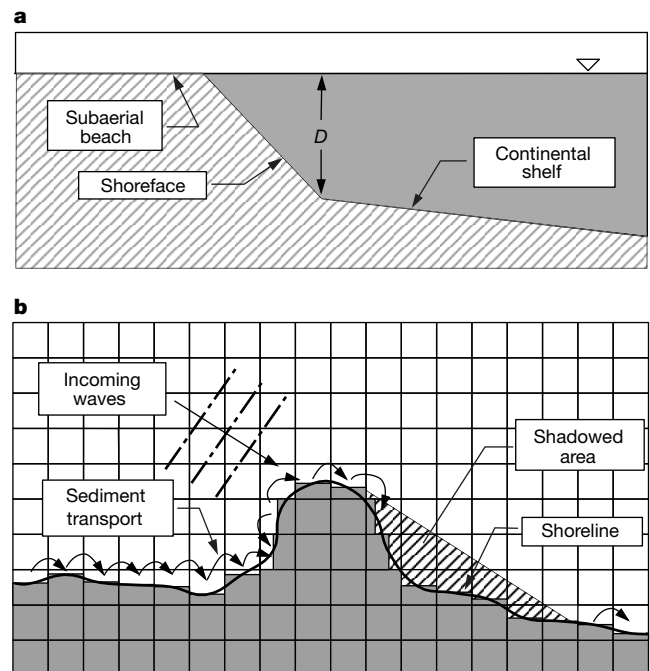


Figure 2 Schematic illustration of the model algorithm. This algorithm, described in Methods, allows the treatment of an arbitrarily complex shoreline. **a**, Profile in the cross-shore direction showing D , the depth at which cross-shore transport is considered negligible, defined as the intersection of the continental shelf with the shoreface, with an imposed minimum depth. The inverted triangle indicates sea level. **b**, Plan view showing the direction of sediment transport between discretized shoreline cells for waves approaching at a given angle. Also shown is the region of the shoreline 'shadowed' from the incoming waves; no sediment is transported between cells in this area.

climate along the flanks also becomes low angle; the evolution of large-scale features inhibits the growth of smaller-scale perturbations (Fig. 3c). Simulations with a slightly asymmetric wave-angle distribution exhibit both forms of behaviour (Fig. 3d).

With greater wave asymmetry and a strong predominance of high-angle waves, large spits extend offshore, and they shelter a downdrift region from waves coming from the dominant direction. Perturbations will then grow in this shadowed area that migrate in a direction opposite to the regional alongshore drift and eventually merge with the larger spit (Fig. 3e).

In simulations with highly asymmetric wave distributions, simulated features resemble a class of shoreline phenomena given various names, including 'cusped spits'^{2,6}, 'sandwaves'^{7–11}, 'ords'¹², and 'humps'¹³. On some open ocean coasts, these can be subtle, migrating, spatially correlated zones of accretion alternating with erosion 'hot spots'^{7,8,10,13}. Pronounced spit- or wave-shaped shoreline features exist in many elongate water bodies, such as the Great Lakes¹⁴, the Sea of Azov² (Fig. 4a) and some back-barrier lagoons^{2,6}. In these water bodies, most of the large waves are generated across the larger fetch along the long axis; even isotropically distributed winds will result in a dominance of high-angle waves.

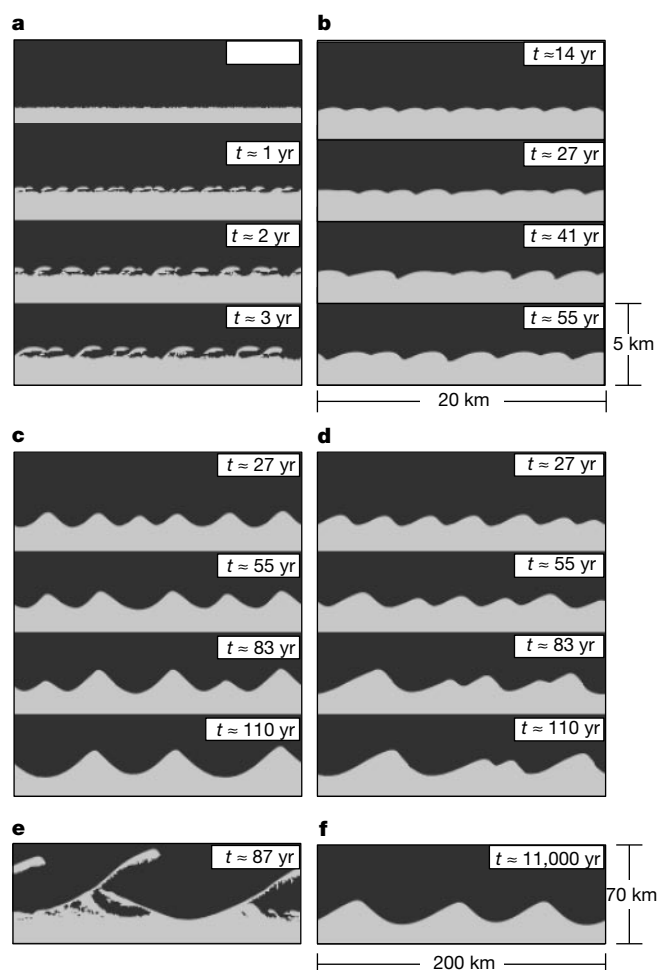


Figure 3 Plan views of the model domain, showing the shoreline evolution. Cell width is 100 m (except where noted). **a**, Results using a single angle (55°) between deep-water wave crests and the global shoreline orientation. **b**, Results when the distribution of deep-water wave angles is weighted towards high angles and towards waves approaching from the left. In this simulation, $\alpha = 0.8$ and $\beta = 0.8$ (see Methods). **c**, Results with a wave distribution weighted toward high angles ($\alpha = 0.6$), but with waves symmetrically distributed from left and right ($\beta = 0.5$). **d**, Slight asymmetry ($\alpha = 0.7$, $\beta = 0.6$). **e**, Longer-term evolution of a cusped spit ($\alpha = 0.6$, $\beta = 0.7$). **f**, Results for $\alpha = 0.7$ and $\beta = 0.57$, with cell width representing 1 km.

With a symmetric or slightly asymmetric wave distribution in the model, simulated features resemble 'cusped forelands'^{15,16} or 'v-notches'¹⁷ found in nature. After a simulated time of the order of 10^4 years (with a constant offshore wave height of 2 m and wave period of 5 s), the cusped pattern develops a length scale of the order of 100 km (Fig. 3f), commensurate with some of the largest cusped forelands found in nature.

The capes of the North Carolina coast of North America provide the most dramatic example of this phenomenon (Fig. 4b). Many hypotheses for the origin of these capes have been suggested, most involving external templates such as ancient river deltas^{18,19}. However, estimates of the modern wave climate off the Outer Banks of North Carolina suggest that a predominance of the wave energy comes from high angles (Fig. 4b)²⁰. Although these estimates do not conclusively establish the regional wave climate, and do not necessarily reflect previous wave conditions, they are consistent with the hypothesis that the large-scale pattern of this coastline has emerged spontaneously, driven by the instability inherent in the

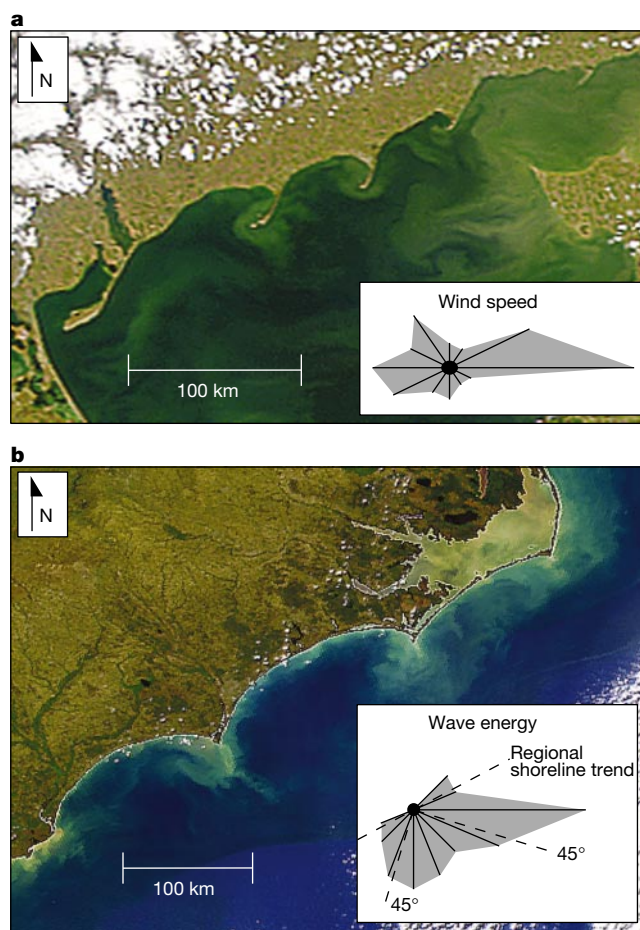


Figure 4 Satellite images showing naturally occurring large-scale shoreline features. **a**, The Sea of Azov, Ukraine, with (inset) directionally binned 'rose' of wind speed showing a dominance of winds from the east. Wind data (from NOAA National Climate Data Centre, Asheville, North Carolina) from 1991–2001 at meteorological station 34712, in Mariupol', on the north shore of the Sea of Azov. **b**, The Carolina coast, USA, with (inset) directionally binned WIS hindcast of wave energy (estimated from meteorological data, solid lines) plotted against the shoreline trend (defined by the backs of the bays, dashed line)²⁰. The relatively large amount of energy approaching from directions between the shoreline trend and 45° to it indicates the predominance of waves with high angles between wave crests and the shoreline trend. Data represent the period 1976–95, at station 52 in 22 m water depth off Cape Hatteras. Neighbouring stations in 27–35 m water depth show a similar distribution. Satellite images courtesy of the SeaWiFS Project NASA/GSFC and ORBIMAGE.

relationship between alongshore sediment transport and local shoreline orientation.

Sea level has remained near its current high elevation (relative to the elevations during Pleistocene glaciations) only of the order of 10^3 years, rather than the 10^4 years required in the model for 100-kilometre-scale features to develop. However, the Carolina capes could have continued to develop over several sea-level high stands; the rapidly rising sea level early in interglacial periods could have encountered abandoned variegated shorelines that evolved during a previous high stand. (Preserved shorelines on the North Carolina coastal plain suggest cape development during Pleistocene times²¹.) Our initial model does not explicitly include sea-level rise or the dynamics of barrier islands. However, during the relatively slow sea-level rise of recent millennia, the development of large-scale shoreline shapes could have been superimposed on the general trend of shoreline regression.

The assumptions used to develop the numerical model are specifically applicable to large spatial and temporal scales. The magnitude of these scales depends on wave height and period (which influence the scale of the alongshore current and the depth to which transport is active). On open-ocean coasts, the model assumptions do not hold for shoreline changes on scales smaller than kilometres and decades. However, during storms producing locally high-angle waves, the fundamental instability would still operate. The resulting spatially isolated zones of erosion could make a shoreline's response to storms more complex than might naively be expected. Some of the transient erosion 'hot spots' that occur along some shorelines after large storms^{22,23} could possibly be related to this instability. Natural shorelines include complicated cross-shore and alongshore bathymetry, and other processes that can produce localized zones of erosion, such as rip currents. However, the recently documented patterns of erosion and accretion exhibit alongshore scales including that of several kilometres²², and a power-law scaling indicating changes on all spatial scales over a broad range²⁴. These observations are inconsistent with a rip-current origin, but this behaviour occurring across a range of scales could be explained by the basic instability in alongshore sediment transport.

This simplified, initial numerical model is limited in several other ways. It applies not to predominantly rocky coasts, but only to shorelines that remain covered with at least a veneer of mobile sediment, even in areas where the shoreline is eroding. If weathering of intermittently exposed rocks on the nearshore bed (the 'shoreface') does not occur as rapidly as sediment is removed, or if the eroding shoreface rocks are composed partly of fine-grained material that does not remain in the nearshore system, the resulting geometry of the shapes will be affected. But none of these processes would counteract the fundamental instability being investigated.

We have also intentionally omitted several other natural phenomena, including: (1) alongshore inhomogeneities, such as variations in grain size and lithology; (2) wave diffraction; (3) the convergence and divergence of wave rays that tends to concentrate wave energy on headlands and away from bays; and (4) the residual motion that arises when tidal currents pass a headland. This last process can transport sediment off a cape onto a subaqueous shoal, acting as a sediment sink that can influence cape dynamics and morphology²⁵.

In this initial numerical exploration, our goal has not been to reproduce any natural system in detail, or to create a maximally accurate model, but rather to investigate the range of coastal landforms and behaviours that could result from the fundamental instability in alongshore sediment transport that occurs in regions with predominantly high-angle waves. The numerical model indicates that this instability could be responsible for the self-organization of a wide variety of well documented but poorly understood shoreline phenomena. □

Methods

Sediment transport

The flux of alongshore momentum that waves carry towards the shore, termed the alongshore 'thrust'²⁶, drives an alongshore current and sediment transport. If the relative angle between the waves and the shoreline ($\phi_b - \theta$) approaches zero, the component of the wave-momentum flux in the alongshore direction approaches zero. If the relative wave angle approaches 90° , then the rate at which wave momentum and energy are transported towards the shore approaches zero. Between these extremes, the thrust reaches a maximum. A well-known semi-empirical formula for Q_s can be expressed in terms of this thrust multiplied by the breaking-wave celerity^{27,28}. Assuming both refraction over shore-parallel contours and no dissipation of wave energy, the thrust is constant outside the surf zone²⁸, so that this formula can be written either in terms of breaking-wave properties or in terms of deep-water quantities: $Q_s = KH_b^{3/2} \sin(\phi_b - \theta) \cos(\phi_b - \theta) = KH_b^{1/2} H_b^2 \sin(\phi_b - \theta) \cos(\phi_b - \theta) \approx K_2 H_0^{12/5} \sin(\phi_b - \theta) \cos^{6/5}(\phi_b - \theta)$, where H is the wave height, K and K_2 are constants and the subscripts b and 0 refer to breaking- and deep-water values, respectively. The deep-water form predicts a maximum in sediment transport for offshore waves approaching the local shoreline at an angle of approximately 42° . An analysis²⁹ of surf-zone processes predicts a very similar curve, with Q_s maximized for a relative offshore wave angle slightly greater than 45° . (In the more familiar breaking-wave formula, with H_b held constant, sediment transport is maximized for $\phi_b - \theta = 45^\circ$. However, with H_0 and ϕ_0 held constant, higher values of $\phi_b - \theta$ are associated with greater nearshore refraction, which tends to reduce H_b . As an example, if $H_0 = 2$ m, wave period is 10 s, and assuming shore-parallel contours, sediment transport is maximized for $\phi_b - \theta \approx 10^\circ$.)

Model

The model divides the plan-view domain into cells that are filled with a fractional amount of sediment, F_i . Cells with $F_i = 1$ form the subaerial coast, and cells adjacent to these with $0 < F_i < 1$ constitute the shoreline, with F_i representing the fractional position of the shoreline within the cell. For each boundary between shoreline cells, the sign of the angle between deep-water wave crests and the line connecting the shoreline positions in adjacent shoreline cells determines the direction of alongshore sediment transport between the neighbouring cells. To avoid discretization artefacts (for the case of high-angle waves), we use a scheme that combines upwind and central finite-difference techniques: if the magnitude of the local relative wave angle—defined using the position of the shoreline in cells bounding the cell sediment transport is from—is less than (greater than) that which maximizes the transport, we use a central (upwind) finite-difference technique to determine the sediment flux. When moving from a boundary with central differencing to one with upwind differencing, sediment transport for the latter boundary is set to that occurring for $\phi_b - \theta = 45^\circ$, with the local H_b . Sediment is not transported out of cells that are 'shadowed' from incoming waves by other portions of the shoreline (Fig. 2b). Sediment transport into a shadowed region is calculated using an upwind finite-difference.

Sediment flux is determined using $Q_s = KH_b^{3/2} \sin(\phi_b - \theta) \cos(\phi_b - \theta)$, where $K = 6.4 \times 10^3 \text{ m}^3 \text{ d}^{-1}$ is used for all simulations¹. (Deep-water wave heights and angles are converted to breaking-wave quantities using linear-wave theory for shoaling and refraction over shore-parallel contours, and assuming depth-limited breaking.) Fractional sediment contents of shoreline cells are adjusted according to the net sediment flux into or out of the cell, $\Delta F_i = Q_{s,\text{net}} \Delta t / (W^2 D_i)$, where W is the horizontal cell width and D_i represents the depth to which profile accumulation or erosion extends (below which cross-shore transport is considered negligible). D_i is defined in the model as the intersection of the continental shelf (which has a slope $S_{cs} = 0.001$, perpendicular to the global shoreline trend) with a linear nearshore profile (representing the 'shoreface', which has a slope $S_{sf} = 0.01$, extending from the subaerial portion of the beach in a direction perpendicular to the local shoreline orientation). Imposing a minimum depth, $D_{\min}$, represents the erosion of the shoreface by wave action where the shore moves landward.

In each simulation, the initial conditions consist of a straight shoreline (at an average of 9,000 m seaward of the intersection of the projection of the continental shelf slope and sea level), plus uncorrelated random, white-noise perturbations to the shoreline position in each shoreline cell, with an amplitude of one cell width. When deep-water wave angles are varied during model runs, each simulated day, (1) a random number between 0 and 1, X , is chosen, and the deep-water wave angle is determined by $\phi_0 = (4\pi/9)X^\alpha$, where α is a set parameter; and (2) a different random number between 0 and 1, Y , is chosen, and sign of the wave angle is positive if $Y > \beta$ and negative if $Y < \beta$ (positive is from the left, negative from the right), where β is a set parameter between 0 and 1.

Received 9 March; accepted 11 September 2001.

1. Komar, P. D. *Beach Processes and Sedimentation* (Simon & Schuster, Upper Saddle River, New Jersey, 1998).
2. Zenkovitch, V. P. On the genesis of cusped spits along lagoon shores. *J. Geol.* **67**, 269–277 (1959).
3. Wang, J. D. & Le Mehaute, B. in *Proc. 17th Coastal Eng. Conf.* 1295–1305 (American Society of Civil Engineers, New York, 1980).
4. Hanson, H. & Kraus, N. C. *GENESIS: Generalized Model for Simulating Shoreline Change, Report 1: Technical Reference* (US Army Eng. Waterways Experiment Station, Coastal Eng. Res. Cent., Vicksburg, Mississippi, 1989).
5. Le Mehaute, B. & Saldade, M. in *Proc. 16th Coastal Eng. Conf.* 1163–1179 (American Society of Civil Engineers, New York, 1979).
6. Fisher, R. L. Cusped spits of Saint Lawrence Island, Alaska. *J. Geol.* **63**, 133–142 (1955).
7. Bruun, P. in *Proc. 5th Conf. on Coastal Eng.* 269–295 (American Society of Civil Engineers, New York, 1954).
8. Bakker, W. T. A mathematical theory about sandwaves and its applications on the Dutch Wadden Isle of Vlieland. *Shore Beach* **36**, 4–14 (1968).

9. Sonu, C. J. in *Proc. 11th Conf. on Coastal Eng.* Vol. 1, 373–400 (American Society of Civil Engineers, New York, 1969).
10. Verhagen, H. J. Sand waves along the Dutch coast. *Coastal Eng.* **13**, 129–147 (1989).
11. Thevenot, M. M. & Kraus, N. C. Longshore and sand waves at Southampton Beach, New York: observation and numerical simulation of their movement. *Mar. Geol.* **126**, 249–269 (1995).
12. Pringle, A. W. Holderness coast erosion and the significance of ords. *Earth Surf. Processes Landforms* **10**, 107–124 (1985).
13. Inman, D. L. Accretion and erosion waves on beaches. *Shore Beach* **55**, 61–66 (1987).
14. Stewart, C. J. & Davidson-Arnott, R. G. D. Morphology, formation and migration of longshore sandwaves; Long Point, Lake Erie, Canada. *Mar. Geol.* **81**, 63–77 (1988).
15. Gulliver, F. P. Cuspate forelands. *Geol. Soc. Am. Bull.* **7**, 399–422 (1896).
16. Wilson, A. W. G. Cuspate forelands along the Bay of Quinte. *J. Geol.* **12**, 106–132 (1904).
17. Gilbert, G. K. The topographic features of lake shores. *US Geol. Surv. Ann. Rep.* **5**, 69–123 (1885).
18. Hoyt, J. H. & Henry, V. J. Origin of capes and shoals along the southeastern coast of the United States. *Geol. Soc. Am. Bull.* **82**, 59–66 (1971).
19. White, W. A. Drainage symmetry and the Carolina Capes. *Geol. Soc. Am. Bull.* **77**, 223–240 (1966).
20. US Army Corps of Engineers *Wave Information Study* at (<http://bigfoot.wes.army.mil/u003.html>) (2001).
21. Moslow, T. F. & Heron, S. D. Holocene depositional history of a microtidal cusate foreland cape: Cape Lookout, North Carolina. *Mar. Geol.* **41**, 251–270 (1981).
22. List, J. & Farris, A. F. in *Coastal Sediments '99* (eds Drause, N. C. & McLough, W. G.) 1324–1338 (American Society of Civil Engineers, Long Island, 1999).
23. Stockdon, H. F., Holman, R. A. & Sallenger, A. H. Longshore variability of the coastal response to extreme storms. *Eos* **81**, F673 (2000).
24. Tebbens, S. F. & Nelson, E. Wavelet analysis of shoreline change at Cape Hatteras National Seashore. *Eos* **81**, F562 (2000).
25. McNinch, J. E. & Luettich, R. A. Physical processes around a cusate foreland headland: implications to the evolution and long-term maintenance of a cape-associated shoal. *Continental Shelf Res.* **20**, 2367–2389 (2000).
26. Longuet-Higgins, M. S. Longshore currents generated by obliquely incident waves, 1. *J. Geophys. Res.* **75**, 6778–6789 (1970).
27. Komar, P. D. & Inman, D. L. Longshore sand transport on beaches. *J. Geophys. Res.* **75**, 5914–5927 (1970).
28. Longuet-Higgins, M. S. in *Waves on Beaches and Resulting Sediment Transport* (ed. Meyer, R. E.) 373–400 (Academic, New York, 1972).
29. Deigaard, R., Fredsoe, J. & Hedegaard, I. B. Mathematical model for littoral drift. *J. Waterway Port Coast. Ocean Eng.* **112**, 351–369 (1988).

Acknowledgements

This work was supported by the Andrew W. Mellon Foundation.

Correspondence and requests for materials should be addressed to A.A. (e-mail: andrew.ashton@duke.edu).

Female sticklebacks count alleles in a strategy of sexual selection explaining MHC polymorphism

Thorsten B. H. Reusch^{*,†}, Michael A. Häberli^{*,†}, Peter B. Aeschlimann^{*,†} & Manfred Milinski^{*}

^{*} Max Planck Institute for Limnology, Department of Evolutionary Ecology, Postfach 165, 24306 Plön, Germany

[†] These authors contributed equally to this work

The origin and maintenance of polymorphism in major histocompatibility complex (MHC) genes in natural populations is still unresolved¹. Sexual selection, frequency-dependent selection by parasites and pathogens, and heterozygote advantage have been suggested to explain the maintenance of high allele diversity at MHC genes^{2–4}. Here we argue that there are two (non-exclusive) strategies for MHC-related sexual selection, representing solutions to two different problems: inbreeding avoidance and parasite resistance. In species prone to inadvertent inbreeding, partners should prefer dissimilar MHC genotypes to similar ones. But if the goal is to maximize the resistance of offspring towards potential infections, the choosing sex should prefer mates with a higher diversity of MHC alleles. This latter strategy should apply when there are several MHC loci, as is the case in most vertebrates^{2,5}. We tested the relative importance of an ‘allele count-

ing’ strategy compared to a disassortative mating strategy using wild-caught three-spined sticklebacks (*Gasterosteus aculeatus*) from an interconnected system of lakes. Here we show that gravid female fish preferred the odour of males with a large number of MHC class-IIb alleles to that of males with fewer alleles. Females did not prefer male genotypes dissimilar to their own.

Sexual selection, the preference of certain mating partners over others, is ubiquitous among animals^{6,7}. The choosing sex may be able to increase the attractiveness of offspring, or gain direct benefits such as parental care⁸. Another function of mate choice is to increase the fitness of offspring by either choosing ‘good genes’, or avoiding incompatible ‘bad genes’—for example in matings with close kin. Particularly suited for testing the idea of choosiness with respect to genes is the MHC, a multigene family that is important in controlling the vertebrate immune system by presentation of self and foreign peptides to T cells². MHC alleles confer specific resistance against pathogens and parasites. Therefore, mate choice should increase the fitness of offspring by maximizing the heterozygosity at MHC loci^{1,4,9,10}, allowing a wider spectrum of pathogens to be recognized during early infection.

The focus of previous studies was disassortative mating, that is, the preference of dissimilar males or females as a mating partner^{11–13}, as a means of inbreeding avoidance, or in order to increase the heterozygosity at MHC loci^{3,14}. In the context of sexual selection, it has not been taken into account that most vertebrates possess several MHC loci⁵. As a result, there are many possible combinations of alleles at different loci². The chances of choosing a partner with identical MHC haplotypes become very unlikely because the combination of alleles at multiple loci renders the expected likelihood of existing MHC genotypes very small, even under linkage disequilibrium. A mechanism of sexual selection focusing on the distinction between similar and dissimilar MHC genotypes becomes inefficient when increasing parasite resistance is important. Females should rather choose partners that maximize the number of different MHC alleles in their offspring³.

We studied populations of the three-spined stickleback (*Gasterosteus aculeatus*) where we identified high MHC diversity for partial sequences of MHC class-IIb loci, coding for the peptide-binding region. At an estimated six loci¹⁵, we identified 24 distinct sequences in only eight fish from a system of interconnected populations, and many more alleles were identified on the basis of single-strand conformation polymorphisms (SSCP) in a total of 144 fish. We also observed marked differences in the number of different alleles per individual fish, varying between two and eight detectable alleles across all loci (Fig. 1). For one location (Schönsee) we calculated that the chance a female has of mating with an MHC-identical male is only 1% if she mates at random. But the probability of choosing a mate with fewer alleles is 46% under random expectations (see Methods). Therefore, we predicted that females

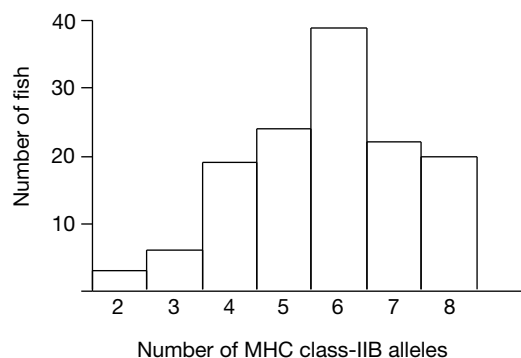


Figure 1 Frequency distribution of the number of MHC class-IIb alleles (peptide-binding region) detectable by SSCP in 144 fish from one population (Schönsee). The mean number of MHC alleles $\pm$ s.e. was 5.8 ± 0.13 .

would choose according to the number of different MHC alleles in males rather than according to similarity/dissimilarity of genotype.

MHC genes influence individual body odour^{16–19}, which is used as a cue for dissimilarity in the mate choice of mice¹¹ and humans¹³. Therefore, we assessed MHC-related mating preference through odour-related behaviour of gravid fish in a flow chamber.

We tested whether females preferred the odour of males with many different MHC class-IIb alleles to the odour of males with few MHC alleles. We presented gravid females with water for the tanks of two males that differed in the number of MHC class-IIb alleles. Males with 'few' alleles had 3–6 alleles, whereas those with 'many' displayed 6–8 alleles. In experimental combinations, male allele number differed by an average of 2.3. We found that choosing females preferred males with a higher number of alleles over those males with few alleles (paired *t*-test: *t* = 2.39, *n* = 29, *P* = 0.023; Fig. 2a). Since inbreeding avoidance can be ruled out as a mechanism of mate choice (see below), the results support a parasite-driven mechanism¹⁴ that maintains the observed diversity of MHC alleles.

In a second experiment, we tested for disassortative mating. Accordingly, we examined whether females prefer the odour of males that differ in their MHC alleles from the female to that of males that share all alleles with the female, with males having the same number of MHC class-IIb alleles as the female. Female sticklebacks did not prefer the odour of MHC class-IIb dissimilar males (sharing as few alleles as possible with the female) to the odour of MHC class-IIb identical males (*t* = 0.34, *n* = 21, *P* = 0.738, statistical power = 0.802, effect size taken from previous experiment; Fig. 2b). These results indicate that disassortative mating is less important than the strategy of female allele counting in a natural population of sticklebacks.

We can exclude the possibility that preferred males were less related to the females than unpreferred males. Over both experiments, we calculated pairwise relatedness coefficients among all pairs of experimental fish on the basis of microsatellite polymorphism. There was no significant correlation among preference time and individual pairwise relatedness (*n* = 46, correlation coefficient, *r*² = 0.04, *P* = 0.19).

Modelling studies^{3,20} indicate that there is an optimal number of MHC alleles, since too many different alleles increase the risk that self-peptides will mount harmful autoimmune responses. Accordingly, we find that females with 7 and 8 alleles tend to be less choosy of males with more alleles than are females with 3–6 MHC class-IIb alleles (analysis of variance, ANOVA, among all categories of female

fish possessing 3–8 different alleles, response variable 'time spent on side of male with more alleles', *P* = 0.037; subsequent contrast 3–6 alleles versus 7 and 8 alleles, *F*_{1,27} = 3.2, *P* = 0.08).

Choosing many rather than few alleles may be a strategy to maximize the general genomic heterozygosity of offspring and not the diversity specifically at MHC loci. However, the number of detectable MHC class-IIb sequences per individual was not correlated with genome diversity, measured at seven microsatellite loci (multi-locus heterozygosity: *r*² = 0.0004, *P* = 0.8; mean *d*², a measure of microsatellite heterozygosity²¹: *r*² = 0.016, *P* = 0.3; both *n* = 124).

The experimental set-up enabled us to test odour preference, as a precursor of mate choice^{13,16,17,19}. There are strong indications that female sticklebacks were indeed choosing mating partners in the experimental runs. Most females spawned spontaneously immediately after the experiments, even in the absence of a male.

The observed MHC-related mating patterns may be direct or indirect. An alternative, but less likely, interpretation of indirect MHC-correlated preference is that fish with high overall MHC class-IIb heterozygosity were less infected with parasites, and that this can be smelled by females. However, experimental males were not infected with parasites that are known to have a pronounced impact on their fitness²¹. Furthermore, screening of 263 wild-caught fish revealed no simple correlation between infection status and MHC class-IIb allele number (K. M. Wegner, T.B.H.R. and M. Kalbe, unpublished work).

Which of the two proposed mate-choice strategies that are associated with MHC polymorphism is more important will depend on the natural population structure of the species studied. In sticklebacks, the chances of mature fish mating accidentally with close kin are low²². On the other hand, in species (mouse, rat, human) where avoidance of similar MHC genotypes (that is, disassortative mating) has been found, the social genetic structure introduces the danger of encountering close kin as a mating partner^{11,12}. Inbreeding avoidance through MHC-related choice is facilitated in species with extended parental care, such as mice, since progeny can learn the MHC odour of its kin by familial imprinting²³. In salmon (*Salmo salar*) adults return to their natal spawning grounds and may locally meet close kin. Nevertheless, there is no indication of inbreeding avoidance associated with MHC alleles²⁴. Salmon possess only one MHC class-IIb locus, so no preference for allele number can be expected. Instead, mating patterns are nonrandom with respect to the distance of MHC alleles in terms of their amino-acid sequence. The inclusion of more species with different population structures will allow us to examine the relative proportion of both MHC-correlated mate choice strategies: disassortative mating and allele counting.

We have thus shown that in natural populations of fish, females will increase individual heterozygosity and hence the population-wide polymorphism at MHC loci, by choosing males with many alleles. We suggest that such a strategy is ubiquitous in vertebrates with several MHC class-IIb loci. □

Methods

MHC characterization

In order to characterize the MHC class-IIb genes of *G. aculeatus*, 96 sequences, spanning the complete exon 2 (coding for the peptide-binding region, PBR) were obtained from two populations of three-spined sticklebacks (Schöensee, *n* = 6 individuals; Ascheberg, *n* = 2). On the basis of intron 3-length variation, we and others¹⁵ estimate that there are up to six MHC class-IIb loci in *G. aculeatus*. For cloning, we used all necessary precautions to avoid erroneous overestimations of allele diversity (that is, proof-read polymerase, plasmid preparation as sequencing template). The forward primer was: AAC TCC ACT GAG CTG AAG GAC ATC (ref. 15). The antisense primer (GCA GTA CAA GTA CCA GTC ATG TAC) was a consensus primer placed in intron 3 on the basis of published sequences¹⁵ and sequences from local populations. Sequences, which were obtained at least twice independently, were translated to amino-acid sequences and revealed a large homology to published MHC class-IIb sequences, but nevertheless represented new alleles in all cases. Amino-acid sequence divergences among alleles were between 2 and 29%. Sequences were deposited in GenBank under accession numbers AF395709–730.

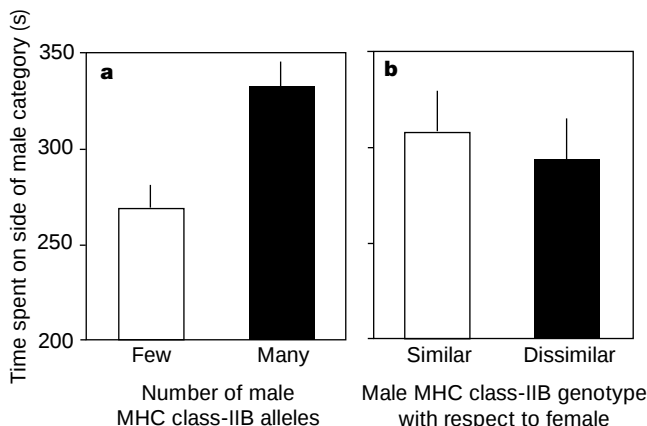


Figure 2 Odour preference of gravid female three-spined stickleback for males in a flow tank. **a**, Mean (+ standard error, s.e.) duration (in seconds) spent on side of male category with either few (2–6) or many (6–8) MHC class-IIb alleles ('many' always having more alleles than 'few'); sample size *n* = 29. **b**, Mean duration (+s.e.) spent on side of male category sharing either all MHC class-IIb alleles (identical to female) or as little as possible (different from female); sample size *n* = 21.

SSCP genotyping

SSCP genotyping was performed using polymerase chain reaction (PCR) with exon 2 internal primers, amplifying a 120-base pair portion of the PBR, and subsequent capillary electrophoresis on an ABI310 automated sequencer²⁵ with modifications. These were the inclusion of a second different reverse primer (GTT GTG CAG ACA GTA AAC CTC CTT C) to increase the number of alleles detectable with SSCP. Both primer pairs together amplified 17/24 (thus 74%) of the PBR that was identified by sequencing in the experimental fish. Isolated sequences cloned into plasmids were subjected to SSCP in order to standardize the fluorescent signals with sequence information. In three cases, complete PBR residues that differed only in 2–4 amino acids cannot be resolved by SSCP, such that our method underestimated the true MHC class-II diversity. In all cases, SSCP signals obtained from PCR of plasmids were also present in genotypes based on fish DNA.

Microsatellite genotyping, heterozygosity and relatedness

All fish of the first experimental block were genotyped for seven polymorphic microsatellites (GenBank accession numbers: AJ010352, -54, -55, -57, -58, -60; ref. 26), representing a total of 72 alleles. Individual heterozygosity was calculated as the number of different alleles per microsatellite locus, and as mean d^2 taking the length difference of alleles into account²⁷. We calculated a correlation between allele number of MHC class-II loci and individual heterozygosity. Relatedness coefficients R (ref. 28) were calculated for all pairs of fish using a program called Relatedness 5.0 (K. F. Goodnight, available from <http://gsoft.smu.edu/gsoft.html>). R coefficients were tested for correlation with female mating behaviour using the time difference of females among preferred/unpreferred male as variable for the female preference.

Permutation procedure

The risk that females will choose an MHC-identical male, or a male with fewer alleles, when mating completely at random was assessed in a permutation procedure. We combined 46 male and 46 female genotypes from Schöhsee 100 times at random, each time counting genotypic similarity and difference in allele number.

Sexual selection experiments

Adult three-spined sticklebacks were seine-netted from an interconnected natural system of large lakes near Plön, Germany, in spring. A bit of a dorsal spine of each fish was cut for MHC class-II and microsatellite genotyping (see above). Thereafter fish were housed in individual tanks (10 litres, continuous exchange of 1 litre per hour, 18 °C, 18 h of light). Males (every second tank) were offered artificial nesting material. Only bright red males with completed nests were used for experiments.

Gravid females were tested in a flow channel²². In the test compartment (25 × 20 cm long, 10 cm water level) each female was offered the choice between the two sides of the current (1.4 cm s⁻¹), which was video-recorded from above. The inlet compartment, which was separated upstream by a net from the test compartment, was divided laterally into halves. A peristaltic precision pump (ISMATEC MV, 100 W, 50 Hz) supplied water (2 × 65 ml min⁻¹) taken from the tanks of each of two males with defined MHC II properties through silicon tubes to the halves of the inlet compartment in the sequence: 2 min neutral water, 5 min 'male' water, 2 min neutral water, 5 min 'male' water after switching sides. The bottles with the stimulus water had been coded by a third person. Each of 50 females was tested only once. Of the 57 males in both experiments, there were three combinations in the disassortative experiment which were tested twice.

Received 3 July; accepted 6 September 2001.

1. Apanius, V., Penn, D., Slev, P. R., Ruff, L. R. & Potts, W. K. The nature of selection on the major histocompatibility complex. *Crit. Rev. Immunol.* **17**, 179–224 (1997).
2. Klein, J. *Natural History of the Major Histocompatibility Complex* 615–620 (Wiley, New York, 1986).
3. Penn, D. & Potts, W. K. The evolution of mating preferences and major histocompatibility complex genes. *Am. Nat.* **153**, 145–164 (1999).
4. Edwards, S. V. & Hedrick, P. W. Evolution and ecology of MHC molecules: from genomics to sexual selection. *Trends Ecol. Evol.* **13**, 305–311 (1998).
5. Kasahara, M. The chromosomal duplication model of the major histocompatibility complex. *Crit. Rev. Immunol.* **167**, 17–32 (1999).
6. Darwin, C. *The Origin of Species by Means of Natural Selection* 88 (Murray, London, 1859).
7. Hamilton, W. D. & Zuk, M. Heritable true fitness and bright birds: a role for parasites? *Science* **218**, 384–387 (1982).
8. Andersson, M. *Sexual Selection* (Princeton Univ. Press, Princeton, New Jersey, 1994).
9. Hamilton, W. D., Axelrod, R. & Tanese, R. Sexual reproduction as an adaptation to resist parasites. *Proc. Natl Acad. Sci. USA* **87**, 3566–3573 (1990).
10. Brown, J. L. A theory of mate choice based on heterozygosity. *Behav. Ecol.* **8**, 60–65 (1997).
11. Potts, W. K., Manning, C. J. & Wakeland, E. K. Mating patterns in seminatural populations of mice influenced by MHC genotype. *Nature* **352**, 619–621 (1991).
12. Ober, C. *et al.* HLA and mate choice in humans. *Am. J. Hum. Genet.* **61**, 497–504 (1997).
13. Wedekind, C., Seebeck, T., Bettens, F. & Paepke, A. J. MHC-dependent mate preferences in humans. *Proc. R. Soc. Lond. B* **260**, 245–249 (1995).
14. Potts, W. K. & Slev, P. R. Pathogen-based models favoring MHC-genetic diversity. *Immunol. Rev.* **143**, 181–197 (1995).
15. Sato, A., Figueroa, F., O'Huigin, C., Steck, N. & Klein, J. Cloning of major histocompatibility complex (MHC)-genes from threespine stickleback, *Gasterosteus aculeatus*. *Mol. Mar. Biol. Biotechnol.* **7**, 221–231 (1998).
16. Yamazaki, K. *et al.* Control of mating preferences in mice by genes in the major histocompatibility complex. *J. Exp. Med.* **144**, 1324–1335 (1976).
17. Brown, R. E., Roser, B. & Singh, P. B. Class I and class II regions of the major histocompatibility complex both contribute to individual odors in congenic inbred strains of rats. *Behav. Genet.* **19**, 659–674 (1989).

18. Singh, P. B., Brown, R. E. & Roser, B. MHC antigens in urine as olfactory recognition cues. *Nature* **327**, 161–164 (1987).
19. Penn, D. & Potts, W. K. How do major histocompatibility complex genes influence odor and mating preferences? *Adv. Immunol.* **69**, 411–435 (1998).
20. Nowak, M. A., Tarczy-Hornoch, K. & Austyn, J. M. The optimal number of major histocompatibility complex molecules in an individual. *Proc. Natl Acad. Sci. USA* **89**, 10896–10899 (1992).
21. Wootton, R. J. in *The Biology of Sticklebacks* (ed. Wootton, R. J.) 75–98 (Academic, New York, 1976).
22. Steck, N., Wedekind, C. & Milinski, M. No sibling odor preference in juvenile three-spined sticklebacks. *Behav. Ecol.* **10**, 493–497 (1999).
23. Yamazaki, K. *et al.* Familial imprinting determines H-2 selective mating preferences. *Science* **240**, 1331–1332 (1988).
24. Landry, C., Garant, D., Duchesne, P. & Bernatchez, L. Good genes as heterozygosity: the major histocompatibility complex and mate choice in Atlantic salmon (*Salmo salar*). *Proc. R. Soc. Lond. B* **268**, 1279–1285 (2001).
25. Binz, T., Reusch, T. B. H., Wedekind, C. & Milinski, M. SSCP analysis of Mhc class II B genes in the threespine stickleback. *J. Fish Biol.* **58**, 887–890 (2001).
26. Llargiàder, C. R., Fries, V., Kobler, B. & Bakker, C. M. Isolation and characterization of microsatellite loci from the three-spine stickleback, *Gasterosteus aculeatus*. *Mol. Ecol.* **8**, 342–344 (1999).
27. Coulson, T. N. *et al.* Microsatellites measure inbreeding depression and heterosis in red deer. *Proc. R. Soc. Lond. B* **265**, 489–495 (1998).
28. Queller, D. C. & Goodnight, K. F. Estimating relatedness using genetic markers. *Evolution* **43**, 258–275 (1989).

Acknowledgements

J. Kurtz, Å. Langefors, D. Penn and W. K. Potts improved earlier versions of the manuscript. We thank S. Liedtke for laboratory assistance, H. Deiwick and D. Lemcke for technical support, and G. Augustin for maintaining the aquaria. P.A. and M.H. were supported by the Swiss National Fund.

Correspondence and requests for materials should be addressed to T.B.H.R. (e-mail: reusch@mpil-ploen.mpg.de).

Illusory perceptions of space and time preserve cross-saccadic perceptual continuity

Kielan Yarrow*, Patrick Haggard†, Ron Heal‡, Peter Brown* & John C. Rothwell*

* Sobell Department of Neurophysiology, Institute of Neurology, 8-11 Queen Square, London WC1N 3BG, UK

† Institute of Cognitive Neuroscience, Department of Psychology, UCL, 17 Queen Square, London WC1N 3AR, UK

‡ Centre for Functional Magnetic Resonance Imaging of the Brain, Department of Clinical Neurology, University of Oxford, John Radcliffe Hospital, Oxford OX3 9DU, UK

When voluntary saccadic eye movements are made to a silently ticking clock, observers sometimes think that the second hand takes longer than normal to move to its next position¹. For a short period, the clock appears to have stopped (chronostasis). Here we show that the illusion occurs because the brain extends the percept of the saccadic target backwards in time to just before the onset of the saccade. This occurs every time we move the eyes but it is only perceived when an external time reference alerts us to the phenomenon. The illusion does not seem to depend on the shift of spatial attention that accompanies the saccade. However, if the target is moved unpredictably during the saccade, breaking perception of the target's spatial continuity, then the illusion disappears. We suggest that temporal extension of the target's percept is one of the mechanisms that 'fill in' the perceptual 'gap' during saccadic suppression. The effect is critically linked to perceptual mechanisms that identify a target's spatial stability.

Although most observers have experienced the 'stopped clock' illusion, previous psychophysical experiments that have tested when subjects perceive the time of transient external events relative to saccadic eye movements have yielded contradictory results^{2,3}. A

possible reason for this is that the saccade itself causes changes in temporal perception at around the time of eye movement. We tested whether the perceived duration of chronostasis is affected by the duration of the saccade. Subjects made saccades of either 22° or 55° degrees (lasting on average 72 and 139 ms, respectively) to a numerical counter. The movement of the eyes was used to start the counter (which proceeded at one increment per second) with the exception that the duration of the first number could be varied between 400 and 1,600 ms. Subjects had to state whether the time that they had seen the first digit was more or less than that for the subsequent digits (a constant 1 s). Figure 1 shows that subjects thought they had seen the initial digit for 1 s when their gaze had been on the target for only 880 ms (22° saccade) or 811 ms (55° saccade). Control trials in which the same temporal judgement was made either without moving the eyes, or if the target rather than the eye saccaded into the visual field, gave significantly different values that were very close to the correct value of 1 s. There was an almost exact agreement between the extra time taken for the eyes to move over the longer distance (139 – 72 = 67 ms) and the difference in the time intervals that subjects matched to 1 s (880 – 811 = 69 ms), suggesting that the illusion of chronostasis is linked to the time taken to move the eyes. In fact, subjects seemed to extend the time that they thought they had seen the first target back in time to approximately 50 ms before the start of the eye movement.

Although subjects reported no awareness of the counter changing during their saccades, it is possible that they were able to use this digit shift as a cue to initiate time judgements. This would invalidate the matched times we calculated (measured from the moment the eyes actually reached the counter). However, a control experiment in which the counter was triggered either very early or very late during a large (55°) saccade showed no difference in the duration of chronostasis, despite modifying the period that the digit was actually on screen by 85 ms.

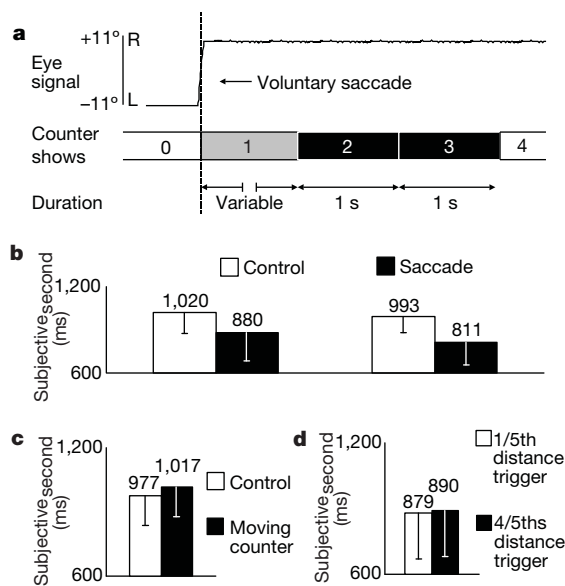


Figure 1 Results for experiment 1. Error bars show standard deviations. **a**, Schematic of experimental method. **b**, Mean time (ms) matched to 1 s for two conditions involving saccades of 55° and 22° (right and left pf figure, respectively) and two control conditions (without saccades) at matched eccentricities. Chronostasis occurs in both experimental conditions compared with controls (55°: $t_{29} = 9.612$, $P < 0.001$; 22°: $t_{29} = 5.608$, $P < 0.001$) and increases linearly in one-to-one correspondence with saccade duration ($t_{29} = 2.553$, $P < 0.05$). **c**, Results for a control experiment where the counter moves to the point of fixation. Chronostasis is not obtained. **d**, Results for a comparison between the standard interval from saccadic onset to counter change and a much longer interval. The duration of chronostasis is unaffected.

The tight coupling of the duration of chronostasis to the duration of the saccade suggests that the effect may be linked to the perceptual gap caused by saccadic suppression and retinal blur that occurs when we move the eyes^{3,4}. However, it is possible that the illusion of chronostasis is not tightly coupled to movement of the eyes per se, but occurs because subjects also shift the locus of their visual attention at around the time their eyes move⁵. This attention shift may act as the reference point to which the target is predated. To test this, subjects were asked either to make the usual saccade to the target or first to shift their attention to the target and then move their eyes. Figure 2a shows that the illusion of chronostasis persisted with a similar magnitude when subjects shifted their attention before moving their eyes. Control trials intermixed with the eye movement trials verified that subjects were successful in shifting the locus of their visual attention⁶. They fixated on a central cross and had to saccade to a target appearing on the right or left of the screen. If they had been told to shift their attention to the correct side before the target appeared, their reaction time was faster than if they had been incorrectly cued (Fig. 2b).

Although chronostasis is linked to voluntary saccades, the coupling is not obligatory—there is at least one condition under which the illusion is not experienced. We designed a third experiment in which the positional stability of the target counter was systematically broken. Subjects made a saccade to the target, but in some trials the computer displaced the target by up to 9° during the time the eyes were moving. Under such conditions, subjects sometimes fail to notice the shift and make an unusually large corrective saccade to fixate the target^{7,8}. Trials were divided into three types: (1) those in which the counter remained stationary throughout; (2) those in which it was moved but the movement was not perceived by the subject; and (3) trials in which target movement was perceived by the subjects. Figure 3 shows the results of this experiment. When there was no target motion, subjects experienced the usual illusion of chronostasis when they made a saccade compared with a control condition with no movement of the eyes. However, if the target was moved and subjects noticed the movement, then no effect was found relative to control. If the shift

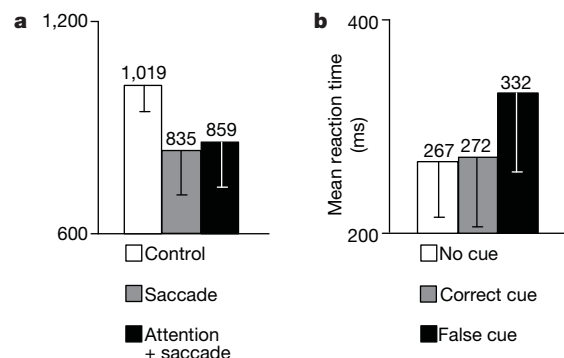


Figure 2 Results for experiment 2. Error bars show standard deviations. **a**, Mean time (ms) matched to 1 s for two conditions involving saccades with/without early, deliberate reorienting of attention, and a control condition. Shifts of attention cannot account for chronostasis because covertly shifting attention early on does not influence the size of the effect. The low value for subjective seconds appears to differ from the results related to saccade duration of experiment 1 (Fig. 1) for a shorter (12°) saccade. However, intersubject variability is high for this task; when data for only those 9 subjects who participated in both studies is considered, the results continue to support a linear effect size scaling with saccade duration. **b**, Mean reaction time (ms) for a two-choice saccade task with no cue to direct attention, a correct cue, or an incorrect cue. Subjects succeeded in reorienting attention, as confirmed by the significantly lower reaction time for the correct cue and no attention conditions relative to the incorrect cue condition (correct cue: $t_{11} = 4.108$, $P < 0.01$; no attention: $t_{11} = 5.367$, $P < 0.001$). Error data (not shown) displayed a similar pattern.

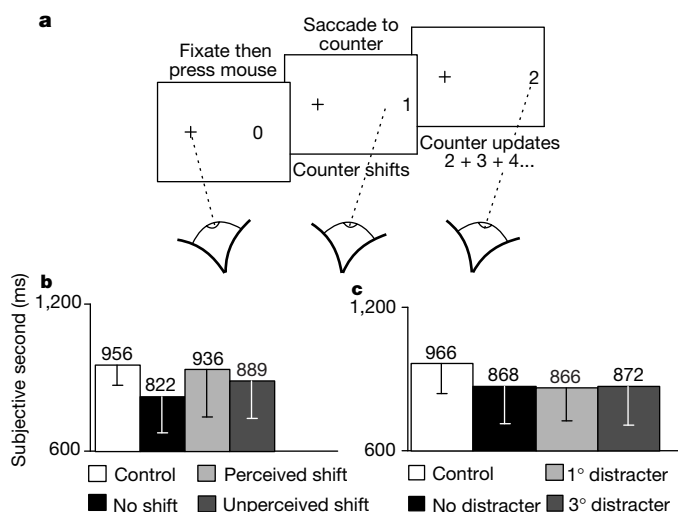


Figure 3 Results for experiments 3 and 4. Error bars show standard deviations. **a**, Schematic of a shift trial (experiment 4). **b**, Mean time (ms) matched to 1 s for four conditions: standard saccade (20°); saccade with detected counter displacement ($\pm 0-9^\circ$); saccade with undetected counter displacement ($\pm 0-9^\circ$); and control. Chronostasis (standard saccade $t_{21} = 4.283$, Bonferroni $P < 0.01$) is eliminated when saccade target stability is noticeably violated and moderated when such a violation goes

unnoticed. **c**, Mean time matched to 1 s (ms) for four conditions: standard saccade (20°); saccade with appearance of a distracter 1° from the target; saccade with appearance of a distracter 3° from the target; and control. Chronostasis is obtained regardless of the distracter (standard: $t = 3.50$, Bonferroni $P < 0.05$; 1° distracter: $t = 3.220$, Bonferroni $P = 0.063$; 3° distracter: $t = 3.724$, Bonferroni $P < 0.05$).

was not perceived, the estimates of the subjects fell somewhere between the control value and the full illusory effect. The effect of moving the target was not due to nonspecific distraction caused by the shift. The full illusion was again obtained in our final experiment, in which distracting stimuli appeared randomly 1° or 3° to the side of the target during the time the eyes were moving, and remained on the screen thereafter (Fig. 3).

Thus, backwards extension of the perception of the target only occurs when subjects perceive that the saccadic target was stationary during the period of extension. We suggest that this link between space and time occurs because of the following. When the saccade shifts the eyes from one stationary viewpoint to another, vision is degraded and it is not possible to say with certainty whether there are any changes in the position of objects during movement. However, if the saccadic target is fixated accurately at the end of the saccade, subjects can assume that it occupied approximately the same place throughout the eye movement (object constancy). Such an assumption may fulfil various functions, having already been proposed in recent theories relating to the problem of space constancy⁹⁻¹¹. As there is no other competing percept (because vision is blurred during the saccade), the assumption of a constant target position is linked to an extended temporal perception of the object as seen at the end of the saccade. If the target jumps, then object constancy is broken, and chronostasis fails to occur. Conscious awareness of a target jump may be linked to the assumption of object constancy, but is unlikely to mirror it precisely. This may explain the partial (nonsignificant) effect for targets that shift without the subject becoming aware of this change.

Perception of the target, rather than being extended back to the time of saccadic onset, predates this by up to 120 ms. Although predating of the target's postsaccadic state to a specific pre-motor event (such as the efferent command) is one possibility, it is notable that the processes underlying both saccadic suppression and space constancy are active over a time period extending beyond the saccade itself⁴. Our obtained constant values are similar to the value of 80 ms obtained for presaccadic shifts in neuron receptive fields within the lateral intraparietal area of monkeys¹². They also fit well with human psychophysical data on presaccadic compression of space (the systematic mislocalization of targets flashed around the time of a saccade) and saccadic suppression, which both precede

saccadic onset by 50 ms or more^{13,14}. It therefore seems probable that presaccadic mechanisms will provide an explanation for the time course of chronostasis.

These data support ideas of conscious experience as an ongoing, often post hoc reconstruction emerging from multiple cognitive systems¹⁵⁻²¹. Our suggestions relating to assumed continuity of target appearance fit well with ideas about object files in the visual attention literature^{22,23}. Here, features of a visual object (colour, form, location, and so on) are bound into a single perceptual unit (the object file) that links representational codes established across diverse cortical regions. We suggest that cross-saccadic perceptual continuity, as described here, may represent a specific case of a more general class of phenomena relating to the continuity of perception across shifting states of sensory input. □

Methods

Experimental design

Subjects sat before a 14-inch colour monitor (60 Hz refresh), with their chin supported. Eye movements were recorded using electro-oculography or with an infrared eye tracker (Microguide 1000 spectacles), and sampled at 200 Hz. Stimuli were black on a white background or vice versa, subtending approximately 0.5°. The experiments were controlled by a personal computer interfaced with a 12-bit A/D card (National Instruments DAQ 1200). Counter change was triggered when the eyes had travelled one fifth of the distance to target. We calculated saccade start/end points automatically using a velocity criterion. Repeated measures designs were used throughout, with condition order counterbalanced. We calculated n for each experiment following a power analysis of initial data sets. Later experiments replicate experiment 1 unless stated otherwise.

Experiment 1

Thirty subjects (18 male, mean age 28.2, standard deviation, s.d. 7.4) completed four conditions: saccades of 55° and 22° and two matched control conditions. In the two saccade conditions, subjects fixated on a cross on one side of the screen, initiated the trial with the depression of a key, and then made a voluntary saccade to a target '0' on the other side of the screen. Eye movement triggered a change of digit to a '1', which remained on the screen for 400–1,600 ms; subsequent digits (2, 3) remained on the screen for 1 s each, culminating in the appearance of a '4'. Subjects indicated whether the time that they saw the '1' for was longer or shorter than that for the subsequent digits. Trials where the first saccade recorded did not exceed 90% of the total distance to target were excluded and immediately repeated. In control trials, subjects fixated a '0' at equivalent eccentricity that changed to become the judged digit (1) 500 ms after the subject's key press. The computer controlled the duration of the first digit by a modified binary search (MOBS)²⁴ procedure that homed in on a single, matched estimate (low boundary 400 ms, high boundary 1,600 ms, initial presentation random 800–1,200 ms, five reversals to terminate). Four estimates

were obtained per condition, then corrected post hoc to match the time that the '1' was on screen after target foveation.

Ten subjects (9 male, mean age 30.5, s.d. 7.8) completed a control experiment. They estimated the duration of the first digit when a counter moved 24° to the point of fixation in 100 ms (six screen refreshes), compared with the usual stationary control. A further control experiment ($n = 10$, 9 male, mean age 31.4, s.d. 7.6) varied the time from saccade onset to the initial counter change by triggering this change either one fifth or four fifths of the way into a 55° saccade (randomly within the same block; separate self-terminating MOBS).

Experiment 2

The data of 12 subjects was included in experiment 2 (10 male, mean age 32.8, s.d. 9.3). In addition to a control, subjects completed two conditions requiring 12° saccades to a counter (as experiment 1) with or without deliberate, prior covert shifting of attention. Every other trial was a reaction time task in which subjects fixated the central target and then made a speeded 12° saccade to the appearance of a target '0' to the left or right. An uninformative cue (an arrow pointing to the left or right near fixation) directed attention before the appearance of the '0' in attention-shift blocks.

Experiment 3

Twenty-two subjects performed experiment 3 (16 male, mean age 30.8, s.d. 7.4). We tested three conditions: a 20° saccade to a stationary counter; a 20° saccade in which the counter shifted ± 0 –9° synchronous with triggering of counter onset; and a control. All eye movement data were obtained within a single block type in which subjects made the standard timing judgment and also indicated whether the counter had moved during the saccade. Presentation was controlled by three randomly interleaved (equally probable) self-terminating MOBS. The first of these controlled target time intervals for the stationary counter trials (as in experiment 1), the latter two controlled the size of the target shift in a hypo- or hypermetric direction (0–9°) according to whether the movement was perceived. This ensured that most of the shift trials were close to the subject's point of shift perception, whether perceived or not. For shift trials, the target time interval was randomly generated in the range 400–1,600 ms. Trials were divided between perceived and unperceived shift conditions post hoc. For all conditions, matched time estimates were generated using logistic regression. Subjects initially completed four experimental blocks and four short control blocks, with a single additional block completed where fitted logistic regression lines exceeded $P = 0.05$.

Experiment 4

Ten subjects participated in experiment 4 (7 male, mean age 29.4, s.d. 7.5). We compared four conditions: a 20° saccade to a stationary counter; an identical saccade with a random, lower-case alphabetic character appearing 1° from the counter (hypo- or hypermetrically) at trigger time; a saccade with the character appearing 3° from the counter; and a control. Data for the first three conditions was obtained within a single block type, using three randomly interleaved and self-terminating MOBS.

Received 6 July; accepted 24 September 2001.

1. Brown, P. & Rothwell, J. C. Illusions of time. *Soc. Neurosci. Abstr.* 27th Annu. Meet. **23**, 1119 (1997).
2. Deubel, H., Irwin, D. E. & Schneider, W. X. in *Current Oculomotor Research: Physiological and Psychological Aspects* (eds Becker, W., Deubel, H. & Mergner, T.) 65–70 (Plenum, New York, 1999).
3. Volkman, F. C. & Moore, R. K. in *Visual Psychophysics and Physiology* (eds Armstrong, J. C., Krauskopf, J. & Wooten, B. R.) 353–361 (Academic, New York, 1978).
4. Ross, J., Concetta Morrone, M., Goldberg, M. E. & Burr, D. C. Changes in visual perception at the time of saccades. *Trends Neurosci.* **24**, 113–121 (2001).
5. Rizzolatti, G., Riggio, L. & Sheliga, B. M. in *Attention and Performance 15: Conscious and Nonconscious Information Processing* (eds Umiltà, C. & Moscovitch, M.) 232–265 (MIT Press, Cambridge, Massachusetts, 1994).
6. Posner, M. I. Orienting of attention. *Q. J. Exp. Psychol.* **32**, 3–25 (1980).
7. Mack, A. An investigation of the relationship between eye and retinal image movement in the perception of movement. *Percept. Psychophys.* **8**, 291–298 (1970).
8. Prablanc, C. & Martin, O. Automatic control during hand reaching at undetected two-dimensional target displacements. *J. Neurophysiol.* **67**, 455–469 (1992).
9. Bridgeman, B., Van der Heijden, A. H. C. & Velichkovsky, B. M. A theory of visual stability across saccadic eye movements. *Behav. Brain Sci.* **17**, 247–292 (1994).
10. Currie, C. B., McConkie, G. W., Carlson-Radvansky, L. A. & Irwin, D. E. The role of the saccade target object in the perception of a visually stable world. *Percept. Psychophys.* **62**, 673–683 (2000).
11. Deubel, H., Bridgeman, B. & Schneider, W. X. Immediate post-saccadic information mediates space constancy. *Vision Res.* **38**, 3147–3159 (1998).
12. Duhamel, J. -R., Colby, C. L. & Goldberg, M. E. The updating of the representation of visual space in parietal cortex by intended eye movements. *Science* **255**, 90–92 (1992).
13. Lappe, M., Awatier, H. & Krekelberg, B. Postsaccadic visual references generate presaccadic compression of space. *Nature* **403**, 892–895 (2000).
14. Diamond, M. R., Ross, J. & Moronne, M. C. Extra-retinal control of saccadic suppression. *J. Neurosci.* **20**, 3449–3455 (2000).
15. Dennett, D. & Kinsbourne, M. Time and the observer. *Behav. Brain Sci.* **15**, 183–247 (1992).
16. Kolars, P. & von Grünau, M. Shape and color in apparent motion. *Vision Res.* **16**, 329–335 (1976).
17. Eagleman, D. M. & Sejnowski, T. J. Motion integration and postdiction in visual awareness. *Science* **287**, 2036–2038 (2000).
18. Geldard, F. A. & Sherrick, C. E. Space, time and touch. *Sci. Am.* **255**, 84–89 (1986).

19. Nishida, S. & Johnston, A. Influence of motion signals on the perceived position of spatial pattern. *Nature* **397**, 610–612 (1999).
20. Libet, B., Wright, E. W., Feinstein, B. & Pearl, D. K. Subjective referral of the timing for a conscious sensory experience. *Brain* **102**, 193–224 (1979).
21. Haggard, P., Aschersleben, G., Gehrke, J. & Prinz, W. in *Attention and Performance 19* (eds Hommel, B. & Prinz, W.) (Oxford Univ. Press, Oxford, in the press).
22. Kahneman, D., Treisman, A. & Gibbs, B. J. The reviewing of object files: object-specific integration of information. *Cogn. Psychol.* **24**, 175–219 (1992).
23. Hommel, B., Müseler, J., Aschersleben, G. & Prinz, W. The theory of event coding (TEC): a framework for perception and action planning. *Behav. Brain Sci.* (in the press).
24. Tyrell, R. A. & Owens, D. A. A rapid technique to assess the resting states of the eyes and other threshold phenomena: the modified binary search (MOBS). *Behav. Res. Methods Instrum. Comput.* **20**, 137–141 (1988).

Acknowledgements

This study was supported by the Medical Research Council. We acknowledge the late P. A. Merton for alerting J.R. and P.B. to the importance of the 'stopped clock' illusion.

Correspondence and requests for materials should be addressed to K.Y. (e-mail: k.yarrow@ion.ucl.ac.uk).

Haemoglobin C protects against clinical *Plasmodium falciparum* malaria

David Modiano^{*†}, Gaia Luoni^{*}, Bienvenu Sodiomon Sirima[‡], Jacques Simporé[§], Federica Verra[†], Amadou Konaté[‡], Elena Rastrelli^{*}, Anna Olivieri^{*}, Carlo Calissano^{*}, Giacomo Maria Paganotti^{*}, Leila D'Urbano^{*}, Issa Sanou^{||}, Alphonse Sawadogol^{||}, Guido Modiano[¶] & Mario Coluzzi^{*†}

^{*} Dipartimento di Scienze di Sanità Pubblica, Sezione di Parassitologia, WHO Collaborating Centre for Malaria Epidemiology and Control; and [†] Istituto Pasteur Fondazione Cenci Bolognietti, University of Rome "La Sapienza", 00185, Rome, Italy

[‡] Centre National de Recherche et Formation sur le Paludisme, Ministère de la Santé; and [§] Centre Medical Saint-Camille; and ^{||} Service de Pédiatrie, Centre Hospitalier National Yalgado Ouédraogo, Ouagadougou, 01 BP 2208, Burkina Faso

[¶] Dipartimento di Biologia, Università "Tor Vergata", 00133, Rome, Italy

Haemoglobin C (HbC; $\beta 6\text{Glu} \rightarrow \text{Lys}$) is common in malarious areas of West Africa, especially in Burkina Faso^{1,2}. Conclusive evidence exists on the protective role against severe malaria of haemoglobin S (HbS; $\beta 6\text{Glu} \rightarrow \text{Val}$) heterozygosity³, whereas conflicting results for the HbC trait have been reported^{4–10} and no epidemiological data exist on the possible role of the HbCC genotype. *In vitro* studies suggested that HbCC erythrocytes fail to support the growth of *P. falciparum*^{11,12} but HbC homozygotes with high *P. falciparum* parasitaemias have been observed¹⁰. Here we show, in a large case–control study performed in Burkina Faso on 4,348 Mossi subjects, that HbC is associated with a 29% reduction in risk of clinical malaria in HbAC heterozygotes ($P = 0.0008$) and of 93% in HbCC homozygotes ($P = 0.0011$). These findings, together with the limited pathology of HbAC and HbCC¹³ compared to the severely disadvantaged HbSS and HbSC genotypes and the low β^S gene frequency in the geographic epicentre of $\beta^{\text{C}1,2,14}$, support the hypothesis that, in the long term and in the absence of malaria control, HbC would replace HbS in central West Africa.

Since hominization the human genome has been under selective pressures for resistance to infectious diseases. For example, West African populations are able to escape the infection altogether, with complete protection from *Plasmodium vivax* achieved through the fixation of a Duffy silent allele (fy)¹⁵. In other cases, polymorphic

Table 1 Absence of age effect on β globin genotype frequencies

Age group (years)	Percentage genotype frequencies $\pm$ standard error*											
	AA		AC		AS		CC		SC		SS	
	Cases	Controls	Cases	Controls	Cases	Controls	Cases	Controls	Cases	Controls	Cases	Controls
0–3	76.7 $\pm$ 2.1	70.8 $\pm$ 5.4	19.2 $\pm$ 2.0	13.9 $\pm$ 4.1	3.3 $\pm$ 0.9	12.5 $\pm$ 3.9	0.25 $\pm$ 0.3	1.4 $\pm$ 1.4	0.5 $\pm$ 0.4	1.4 $\pm$ 1.4	0.0 (395)	0.0 (72)
>3–6	85.2 $\pm$ 2.2	66.7 $\pm$ 3.5	13.6 $\pm$ 2.1	19.4 $\pm$ 2.9	1.2 $\pm$ 0.7	11.1 $\pm$ 2.3	0.0 (257)	1.7 $\pm$ 1.0	0.0 (257)	1.1 $\pm$ 0.8	0.0 (257)	0.0 (180)
>6–10	80.2 $\pm$ 3.6	66.4 $\pm$ 1.4	15.7 $\pm$ 3.3	21.7 $\pm$ 1.2	3.3 $\pm$ 1.6	9.5 $\pm$ 0.9	0.0 (121)	1.4 $\pm$ 0.3	0.8 $\pm$ 0.8	0.9 $\pm$ 0.3	0.0 (121)	0.1 $\pm$ 0.1
>10	82.5 $\pm$ 5.0	66.2 $\pm$ 1.0	12.3 $\pm$ 4.3	22.2 $\pm$ 0.9	5.3 $\pm$ 3.0	9.3 $\pm$ 0.6	0.0 (57)	1.8 $\pm$ 0.3	0.0 (57)	0.4 $\pm$ 0.1	0.0 (57)	0.0 (2082)
P values	0.063	0.88	0.21	0.33	0.23	0.71	0.78	0.79	0.55	0.24	1.0	0.58

* When genotype frequency was zero, the number of individuals examined is indicated between brackets. Cases, malaria patients; controls, healthy subjects; from Burkina Faso, West Africa.

gene mutants decreasing the susceptibility to severe forms of *P. falciparum* malaria have been selected in spite of a heavy segregational load (balanced polymorphisms). Since Haldane's malaria hypothesis¹⁶, epidemiological and in some cases *in vitro* evidence on the protective role of thalassaemias^{17,18}—HbS^{3,19,20}, human leukocyte antigen Bw53 (ref. 3), glucose-6-phosphate dehydrogenase deficiency²¹ and haemoglobin E (HbE)²²—have been obtained. Attempts to evaluate if HbC protects against malaria (both the infection and the disease) have yielded conflicting results. The first studies showed no correlations of HbC heterozygosity with parasitological indicators such as *P. falciparum* parasite rates and densities^{4–7,14}. Clinical studies performed in Nigeria⁸ and Mali^{9,10}, assessing the role of HbAC with respect to the susceptibility to severe forms of the disease, showed contrasting results: the Nigerian⁸ and one study from Mali⁹ indicated lack of protection, and the other Malian study¹⁰ showed evidence for an association between HbAC and protection against severe malaria in the Dogon population. In the same study¹⁰, episodes of non-complicated malaria with high *P. falciparum* parasitaemia were observed in HbCC patients, although *P. falciparum* was previously reported not to proliferate in HbCC erythrocytes *in vitro*^{11,12}.

None of the epidemiological studies mentioned above was designed or had the statistical power to evaluate the possible role of HbC homozygosity in the resistance or susceptibility to severe malaria. Moreover, a report on the possible reduced survival of the HbCC genotype¹⁴, together with the need of extremely large clinical and control samples, did not encourage research in this direction.

In a large case-control study of *P. falciparum* malaria in Burkina Faso, we estimated the genotype and allele frequencies for the β globin gene in 3,513 healthy subjects and in 835 malaria patients recruited at the pediatric ward of the Ouagadougou University Hospital with a clinical picture of severe ($n = 359$) or non-severe ($n = 476$) malaria. The AA, AC and CC genotype frequencies were in Hardy–Weinberg equilibrium ($P > 0.6$) among healthy subjects but not among malaria patients ($P = 0.0005$). This disequilibrium is clearly dependent on the almost complete lack of the CC genotype in the clinical group (1 observed versus 13.8 expected). Among healthy subjects, HbSC and HbSS genotypes, in line with their known short life expectancy, were found at much lower frequencies than those expected by the Hardy–Weinberg equilibrium [23 versus 46.2 for HbSC ($P < 0.001$); 1 versus 9.2 for HbSS ($P < 0.01$)]. No evidence of any age effect on genotype frequencies was observed both in the case ($P = 0.40$) and control ($P = 0.73$) samples (Table 1).

As shown in Table 2, obvious indications of protection against clinical malaria were observed not only—as expected—for the HbAS genotype (odds ratio, OR 0.27, 95% confidence intervals, c.i. 0.17–0.42, $P \ll 0.001$) but also for HbAC (OR 0.71, 95% c.i. 0.58–0.87), $P = 0.0008$ and even more strongly for the HbCC genotype (OR 0.07, 95% c.i. 0.00–0.48, $P = 0.0011$). No differences were observed in HbC between severe and non-severe malaria patients, whereas significantly lower β^S allele ($P = 0.023$) and HbAS genotype ($P = 0.021$) frequencies were recorded among severe malaria patients than in those with non-severe malaria. No

Table 2 β globin genotype and allele frequencies in healthy subjects and in malaria patients from Burkina Faso, West Africa

Sample	n	Relative and (absolute) genotype frequencies						Relative and (absolute) allele frequencies		
		AA	AC	AS	CC	SC	SS	β^A	β^C	β^S
Healthy subjects	3,513	0.6641 (2333)	0.2172 (763)	0.0954 (335)	0.0165 (58)	0.0065 (23)	0.0003 (1)	0.8204 (5764)	0.1284 (902)	0.0512 (360)
Severe malaria	359	0.8078 (290)	0.1755 (63)	0.0111 (4)	0.0028 (1)	0.0028 (1)	0 (0)	0.9011 (647)	0.0919 (66)	0.0070 (5)
Non-complicated malaria	476	0.8004 (381)	0.1555 (74)	0.0399 (19)	0 (0)	0.0042 (2)	0 (0)	0.8981 (855)	0.0798 (76)	0.0221 (21)
Malaria patients (total)	835	0.8036 (671)	0.1641 (137)	0.0275 (23)	0.0012 (1)	0.0036 (3)	0 (0)	0.8994 (1502)	0.0850 (142)	0.0156 (26)
Comparisons		Odds ratio (95% confidence interval) and P values*								
Healthy subjects versus malaria patients		2.07† (1.71–2.50)	0.71‡ (0.58–0.87)	0.27§ (0.17–0.42)	0.07 (0.00–0.48)	n.s.	n.s.	$\ll 0.001$	$\ll 0.001$	$\ll 0.001$
Severe malaria versus non-severe malaria		$\ll 0.001$	0.0008	$\ll 0.001$	0.0011	n.s.	n.s.	n.s.	n.s.	0.023
		n.s.	n.s.	0.27¶ (0.08–0.86)	0.021	n.s.	n.s.	n.s.	n.s.	0.023

n, number of individuals examined; n.s., not significant.

* Although no evidence of any age effect on genotype frequencies was observed (Table 1) and the area of residence (urban or rural) of the case and control groups was almost overlapped, the possible confounding effect of these two factors was calculated by Mantel–Haenszel (M–H) weighted odds ratio (OR; 95% confidence intervals) and maximum likelihood estimate (MLE) of OR (exact 95% c.i.) after stratification by age (0–3 years, >3–6, >6–10, >10) and area of residence (urban or rural).

† M–H weighted OR: 2.15 (c.i. 1.66–2.79), $P \ll 0.001$; MLE of OR: 2.13 (c.i. 1.63–2.79); probability of MLE ≥ 2.13 if population OR = 1.0, $P \ll 0.001$.

‡ M–H weighted OR: 0.74 (c.i. 0.55–0.99), $P = 0.047$; MLE of OR: 0.75 (c.i. 0.56–1.00); probability of MLE ≤ 0.75 if population OR = 1.0, $P = 0.0268$.

§ M–H weighted OR: 0.26 (c.i. 0.15–0.45), $P \ll 0.001$; MLE of OR: 0.23 (c.i. 0.12–0.40); probability of MLE ≤ 0.23 if population OR = 1.0, $P \ll 0.001$.

|| M–H weighted OR: 0.03 (c.i. 0.00–0.73), $P = 0.007$; MLE of OR: 0.07 (c.i. 0.00–0.59); probability of MLE ≤ 0.07 if population OR = 1.0, $P = 0.0028$.

¶ M–H weighted OR: 0.27 (c.i. 0.09–0.82), $P = 0.014$; MLE of OR: 0.28 (c.i. 0.07–0.85); probability of MLE ≤ 0.28 if population OR = 1.0, $P = 0.010$.

differences were recorded in the geometric means of *P. falciparum* parasite densities between malaria patients with HbAA (10,162 parasites μl^{-1}) and HbAC (11,066 parasites μl^{-1}) genotypes, whereas significantly lower values were observed among HbAS patients (1,995 parasites μl^{-1} ; AS versus AA, $P = 0.004$; AS versus AC, $P = 0.013$). In contrast to previous observations from Mali¹⁰, we did not record CC homozygotes among non-severe malaria patients. A very low parasite density (8 parasites μl^{-1}) was observed in the only CC subject with a clinical picture of severe malaria. These observations seem to be consistent with previous *in vitro* data indicating that HbCC erythrocytes fail to support the growth of *P. falciparum*^{11,12}.

Clearly, HbC provides protection against clinical *P. falciparum* malaria in both the heterozygous and homozygous state. The estimated reduction in the relative risk of clinical malaria associated with CC homozygosity (93%) is stronger than that of AC heterozygosity (29%) and similar to that of the HbAS genotype (73%). The estimated protection of HbAC (OR 0.71; c.i. 0.58–0.87) is lower than in ref. 10 comparing severe and non-complicated malaria (OR 0.25; c.i. 0.06–0.86) but the confidence intervals of the odds ratio partially overlap. We recorded no differences in HbAC frequency between severe and non-severe malaria: this discrepancy may to some extent derive from the fact that the Malian study¹⁰ was performed in a hospital of a small town (Bandiagara, with 12,000 inhabitants) whereas this work was carried out in the main hospital of a city of about one million inhabitants (Ouagadougou). Our sample of non-severe malaria patients could be biased in the direction of high severity because the Ouagadougou University Hospital is the last level in the therapeutic itinerary of the patient. This may have resulted in the smoothing—though not the suppression (see the AS genotype frequencies in severe versus non-severe malaria)—of the clinical differences between the groups of severe and non-severe malaria. Given the genotype frequencies among healthy subjects (AC, 0.2172; CC, 0.0165; AS, 0.0954) and their respective protections, it can be estimated that in the Mossi population the proportion of potential clinical malaria cases prevented by HbC (AC + CC), that is, $[(0.29 \times 0.2172) + (0.93 \times 0.0165)] = 7.83\%$, is similar to that prevented by the HbAS genotype ($0.73 \times 0.0954 = 6.96\%$). Given the heavy genetic load of β^S due to the reduced fitness of the HbSS and HbSC genotypes (balanced polymorphism) and the probable absent or limited genetic load of β^C (transient polymorphism) it can safely be hypothesized that in those selective malaria contexts where these two alleles coexist the β^C gene is probably replacing β^S (refs 1, 2, 14). The restricted geographic distribution of the β^C gene and the evidence of its unicentric origin in central West Africa²³, together with the present results showing an obvious protection of HbC against clinical *P. falciparum* malaria, may suggest the recent origin of the β^C mutation. The β^C polymorphism has such a favourable cost/benefit ratio in highly malarious contexts that its relatively recent origin would seem puzzling: β^S and β^E , in spite of less favourable cost/benefit ratios, probably had polycentric origins and much wider diffusions. The most direct explanation would be that the rates of the $\beta^A \rightarrow \beta^S$ mutation (A $\rightarrow$ T single nucleotide substitution, SNS) and of the $\beta^A \rightarrow \beta^E$ (G $\rightarrow$ A SNS) are higher than that of the $\beta^A \rightarrow \beta^C$ (G $\rightarrow$ A SNS). This hypothesis is unlikely, not only with respect to β^E but even more so with respect to β^S , because the average G $\rightarrow$ A SNS rate is markedly higher than that of the other SNS^{24,25}. Unlike HbS and HbE, which protect against malaria in the heterozygous state, HbC seems to protect mainly in the homozygous state: therefore, even considering higher or similar $\beta^A \rightarrow \beta^C$ mutation rates with respect to $\beta^A \rightarrow \beta^S$ and $\beta^A \rightarrow \beta^E$, the chance of β^C being selected (in spite of the initially strong adverse odds due to genetic drift) could be smaller, and the time it needs to reach significant frequencies could be longer, compared to β^S and β^E . Moreover, the fact that β^C protects mainly in the homozygous state may suggest a straightforward explanation for its very localized

occurrence in West Africa: its selective advantage would be proportional to the allele frequency so that homozygous β^C would progressively fade out with increasing distance from the epicentre. The ideal epidemiological context for positive selection of a protective factor acting mainly in homozygosity is one with very high malaria transmission levels, such as sub-Saharan West Africa. A recent work²⁶ indicating a multicentric origin for β^C , as for HbS^{27,28} and HbE²⁹, seems to support the positive selection of this allele in a non-African malarious context as well. □

Methods

Study area and subjects

The study was performed in Ouagadougou, Burkina Faso. The area has a rainy season lasting from June to October, which corresponds to the high malaria transmission season, and a long dry season from November to May. Inoculation rates range from 1 to 10 per person per year in urban areas of Ouagadougou, and from 50 to 200 in the surrounding rural zones. The main malaria vectors are *Anopheles gambiae*, *A. arabiensis* and *A. funestus*. About 95% of the population in the Ouagadougou area belongs to the Mossi ethnic group.

Healthy subjects were recruited in 12 primary schools in the urban area of Ouagadougou in the frame of a large programme for screening of haemoglobinopathies in Burkina Faso carried out by the Saint Camille Health Center of Ouagadougou. Of the total of 3,686 individuals examined, 3,513 belonged to the Mossi ethnic group and were considered in this study to be controls. Of the 3,513 control subjects, 3,056 (87%) were primary school children (aged 6–15), the remainder being either children aged 1–5 years ($n = 163$, 4.6%) or subjects more than 15 years old ($n = 294$, 8.4%). The presence of subjects aged 0–5 years was due to the sporadic request to include younger children in free-of-charge analyses and those aged over 15 years were repeater students, teachers and school personnel. The mean age $\pm$ standard deviation (s.d.) of the control group was 11.3 ± 3.9 years (range 1–40). Malaria patients were recruited during an epidemiological study on severe malaria performed in Burkina Faso³⁰. A total of 835 malaria patients (359 severe and 476 non-severe malaria) belonging to the Mossi ethnic group were considered in the present analysis. The clinical study was carried out at the 158-bed paediatric ward of the Ouagadougou University Hospital. Patients aged 6 months to 15 years were included in the study. The mean age $\pm$ s.d. of the clinical sample [4.4 ± 3.2 (range 0.5–15 years)] was lower than that of the control sample. The majority of malaria patients, 93.7% ($n = 782$), came from urban areas of Ouagadougou; the remainder were from surrounding rural zones. Severe malaria was defined by the presence of *P. falciparum* in the thick blood film associated with at least one of the following conditions: prostration (that is, incapacity of the child to sit without help in the absence of coma), unrousable coma (score between 0 and 2 on the Glasgow modified coma scale), repeated generalized convulsions (more than two episodes in the preceding 24 hours), severe anaemia (haemoglobin $<0.05 \text{ g ml}^{-1}$), hypoglycaemia (blood glucose $<0.4 \text{ mg dl}^{-1}$), pulmonary oedema/respiratory distress, spontaneous bleeding and renal failure (plasma creatinine $>0.03 \text{ mg dl}^{-1}$). Non-severe malaria was defined as a clinical illness characterised by an axillary temperature $>37.5^\circ\text{C}$ associated with a *P. falciparum*-positive blood film. Children with other detectable infections that were responsible for the hospital admission were not included in the study. All patients with a clinical picture of severe malaria were hospitalized whereas the majority (409/476) of non-severe malaria patients were recruited at the outpatients section of the paediatric ward of the Ouagadougou University Hospital; the remaining 67 patients with non-severe malaria were hospitalized because of high fever ($\geq 40^\circ\text{C}$) and/or anaemia (haemoglobin between 0.051 and 0.08 g ml^{-1}), high parasite densities ($>10^5 \mu\text{l}^{-1}$), severe diarrhoea, severe vomiting. On admission or at outpatient examination and after oral informed consent of the parents, a blood sample was drawn for measurement of parasitaemia, blood glucose level, plasma creatinine concentrations, haemoglobin concentrations, haematocrit, complete cell count, humoral response to malaria antigens and genetic analyses. Patients were treated according to WHO guidelines with a complete regimen of drugs that were provided free of charge as part of the study. The study protocol was approved by the Centre National de Lutte contre le Paludisme of the Ministry of Health of Burkina Faso.

Haemoglobin typing

It was performed by cellulose acetate electrophoresis on the sample of healthy subjects and by polymerase chain reaction (PCR) restriction fragment length polymorphism (RFLP) on malaria patients. DNA samples were amplified with the following oligonucleotides: sense 5'-AGGAGCAGGGAGGGCAGGA3', and antisense 5'-TCCAAGGGTAGACCACCAGC3' (NCBI, NG_000007). The 358 base pair (bp) fragment was digested with *Mnl* I (5'-CCTC(N)/3'; 3'-GGAG(N)/5') which allows the discrimination among HbAA (173 bp, 109 bp and 60 bp), Hb SS/CC/SC (173 bp, 109 bp and 76 bp), HbAS/AC (173 bp, 109 bp, 76 bp and 60 bp), HbAE (249 bp, 173 bp, 109 bp and 60 bp) and HbEE (249 bp and 109 bp). A second digestion with *Dde* I (C/TNAG) was performed on those samples with the ambiguous result to further discriminate among HbSS (331 bp), HbCC (201 bp and 130 bp) HbSC (130 bp, 201 bp, and 331 bp), HbAS (130 bp, 201 bp, and 331 bp) and HbAC (130 bp, and 201 bp). Both digestions were carried out for 3 h at 37°C and resolved on 3% agarose gel which allows a good detection of small size differences such as between 76 bp and 60 bp fragments. A total of 89 samples consisting of 64 AA, 17 AC, 4 AS, 2 CC, 1 SC and 1 SS, were typed both with electrophoresis and RFLP and no discordant results were observed.

Statistical analysis

P values of the comparisons were obtained by Yates corrected χ^2 test. Since no evidence of any age effect on genotype frequencies was observed (Table 1) and in view of the almost overlapped geographic origin (urban or rural) of the case and control group, unadjusted odds ratios (ORs) were calculated. Mantel–Haenszel weighted OR (95% confidence interval) and maximum likelihood estimate of OR (95% c.i.) were also calculated after stratification by age (0–3 years, >3–6, >6–10, >10) and area of residence (urban or rural). Student's *t*-test was used for the comparisons of age and parasite densities.

Received 5 September; accepted 25 September 2001.

- Livingstone, F. B. *Abnormal Hemoglobins in Human Populations* (Aldine, Chicago, 1967).
- Cavalli Sforza, L. L. & Bodmer, W. F. *The Genetics of Human Populations* (Freeman, San Francisco, 1971).
- Hill, A. V. *et al.* Common West African HLA antigens are associated with protection from severe malaria. *Nature* **352**, 595–600 (1991).
- Thompson, G. R. Significance of Hemoglobin S and C in Ghana. *Br. Med. J.* **1**, 682–685 (1962).
- Thompson, G. R. Malaria and stress in relation to Hemoglobins S and C. *Br. Med. J.* **2**, 976–978 (1963).
- Ringelmann, B., Hathorn, M. K., Jilly, P., Grant, F. & Parniczky, G. A new look at the protection of hemoglobin AS and AC genotypes against *Plasmodium falciparum* infection: a census tract approach. *Am. J. Hum. Genet.* **28**, 270–279 (1976).
- Molineaux, L. & Gramiccia, G. *The Garki Project. Research on the Epidemiology and Control of Malaria in the Sudan Savanna of West Africa* (World Health Organization, Geneva, 1980).
- Gilles, H. M. *et al.* Glucose-6-phosphate-dehydrogenase deficiency, sickling, and malaria in African children in South Western Nigeria. *Lancet* **1**, 138–140 (1967).
- Guinet, F. *et al.* A comparison of the incidence of severe malaria in Malian children with normal and C-trait hemoglobin profiles. *Acta Trop.* **68**, 175–182 (1997).
- Agarwal, A. *et al.* Hemoglobin C associated with protection from severe malaria in the Dogon of Mali, a West African population with a low prevalence of hemoglobin S. *Blood* **96**, 2358–2363 (2000).
- Friedman, M. J., Roth, E. F., Nagel, R. L. & Trager, W. The role of hemoglobins C, S, and Nbal in the inhibition of malaria parasite development in vitro. *Am. J. Trop. Med. Hyg.* **28**, 777–780 (1979).
- Olson, J. A. & Nagel, R. L. Synchronized cultures of *P. falciparum* in abnormal red cells: the mechanism of the inhibition of growth in HbCC cells. *Blood* **67**, 997–1001 (1986).
- Smith, E. W. & Krevans, J. R. Clinical manifestations of hemoglobin C disorders. *Bull. Johns Hopkins Hosp.* **104**, 17–43 (1959).
- Lapie, D., Richin, C., Pagnier, J., Gentilini, M. & Nagel, R. L. Hemoglobins S and C in Upper Volta. *Hum. Genet.* **65**, 300–302 (1984).
- Miller, L. H., Mason, S. J., Clyde, D. F. & McGinniss, M. H. The resistance factor to *Plasmodium vivax* in blacks. The Duffy-blood-group genotype, FyFy. *N. Engl. J. Med.* **295**, 302–304 (1976).
- Haldane, J. B. S. The rate of mutation of human genes. *Heredity* **35** (suppl.), 267–273 (1949).
- Flint, J. *et al.* High frequencies of alpha-thalassaemia are the result of natural selection by malaria. *Nature* **321**, 744–750 (1986).
- Allen, S. J. *et al.* alpha-thalassaemia protects children against disease caused by other infections as well as malaria. *Proc. Natl Acad. Sci. USA* **94**, 14736–14741 (1997).
- Pasvol, G., Weatherall, D. J. & Wilson, R. J. Cellular mechanism for the protective effect of haemoglobin S against *P. falciparum* malaria. *Nature* **274**, 701–703 (1978).
- Friedman, M. J. Erythrocytic mechanism of sickle cell resistance to malaria. *Proc. Natl Acad. Sci. USA* **75**, 1994–1997 (1978).
- Ruwende, C. *et al.* Natural selection of hemi- and heterozygotes for G6PD deficiency in Africa by resistance to severe malaria. *Nature* **376**, 246–249 (1995).
- Hutagalung, R. *et al.* Influence of hemoglobin E trait on the severity of Falciparum malaria. *J. Infect. Dis.* **179**, 283–286 (1999).
- Trabuchet, G. *et al.* Nucleotide sequence evidence of the unicentric origin of the β^C mutation in Africa. *Hum. Genet.* **87**, 597–601 (1991).
- Modiano, G., Battistuzzi, G. & Motulsky, A. G. Nonrandom patterns of codon usage and of nucleotide substitutions in human alpha- and beta-globin genes: an evolutionary strategy reducing the rate of mutations with drastic effects? *Proc. Natl Acad. Sci. USA* **78**, 1110–1114 (1981).
- Crow, J. F. The origins, patterns and implications of human spontaneous mutation. *Nature Rev. Genet.* **1**, 40–47 (2000).
- Sanchaisuriya, K. *et al.* Molecular characterization of hemoglobin C in Thailand. *Am. J. Hematol.* **67**, 189–193 (2001).
- Pagnier, J. *et al.* Evidence for the multicentric origin of the sickle cell hemoglobin gene in Africa. *Proc. Natl Acad. Sci. USA* **81**, 1771–1773 (1984).
- Kulozik, A. E. *et al.* Geographical survey of beta S-globin gene haplotypes: evidence for an independent Asian origin of the sickle-cell mutation. *Am. J. Hum. Genet.* **39**, 239–244 (1986).
- Antonarakis, S. E. *et al.* Evidence for multiple origins of the beta E-globin gene in Southeast Asia. *Proc. Natl Acad. Sci. USA* **79**, 6608–6611 (1982).
- Modiano, D. *et al.* Severe malaria in Burkina Faso: influence of age and transmission level on clinical presentation. *Am. J. Trop. Med. Hyg.* **59**, 539–542 (1998).

Acknowledgements

We thank all the children in this study, and their parents and teachers, for their understanding and assistance. We thank the paediatric and laboratory staff of the Centre Hospitalier National Yalgado Quédraogo and of the Saint Camille Health Centre of Ouagadougou, Burkina Faso, for technical assistance. This work was based at the Centre National de Recherche et Formation sur le Paludisme of the Ministry of Health of Burkina Faso, co-sponsored by the Italian Cooperation (MAE-DGCS). Financial support was also provided by the European Community and by the Conferenza Episcopale Italiana.

Correspondence and requests for materials should be addressed to D.M. (e-mail: david.modiano@uniroma1.it).

Induction of *gadd45 β* by NF- κ B downregulates pro-apoptotic JNK signalling

Enrico De Smaele, Francesca Zazzeroni, Salvatore Papa, Dung U. Nguyen, Rongguan Jin, Joy Jones, Rong Cong & Guido Franzoso

The Gwen Knapp Center for Lupus and Immunology Research, and The Ben May Institute for Cancer Research, The University of Chicago, 924 East 57th Street, Chicago, Illinois 60637, USA

In addition to coordinating immune and inflammatory responses, NF- κ B/Rel transcription factors control cell survival¹. Normally, NF- κ B dimers are sequestered in the cytoplasm by binding to inhibitory I κ B proteins, and can be activated rapidly by signals that induce the sequential phosphorylation and proteolysis of I κ Bs¹. Activation of NF- κ B antagonizes apoptosis or programmed cell death by numerous triggers, including the ligand engagement of 'death receptors' such as tumour-necrosis factor (TNF) receptor². The anti-apoptotic activity of NF- κ B is also crucial to oncogenesis and to chemo- and radio-resistance in cancer². Cytoprotection by NF- κ B involves the activation of pro-survival genes²; however, its basis remains poorly understood. Here we report that NF- κ B complexes downregulate the c-Jun amino-terminal kinase (JNK) cascade³, thus establishing a link between the NF- κ B and the JNK pathways. This link involves the transcriptional upregulation of *gadd45 β /myd118* (ref. 4), which downregulates JNK signalling induced by the TNF receptor (TNF-R). This NF- κ B-dependent inhibition of the JNK pathway is central to the control of cell death. Our findings define a protective mechanism that is mediated by NF- κ B complexes and establish a role for the persistent activation of JNK in the apoptotic response to TNF- α .

To understand mechanisms controlling TNF-R-induced programmed cell death (PCD)- and NF- κ B-dependent survival, we used the method of 'death trap' screening⁵ in NF- κ B null cells. Complementary DNA expression libraries derived from TNF- α -treated wild-type cells were transfected into NF- κ B/RelA^{-/-} fibroblasts⁶. Apoptosis was induced with TNF- α , and plasmids were recovered from surviving cells. After four cycles of selection, about 35% of the library cDNAs protected RelA null cells from killing by TNF- α . Known inhibitors of TNF-R-triggered apoptosis, including RelA, cFLIP (cellular FLICE inhibitory protein) and dominant-negative forms of FADD⁷ (Fas-associated death domain protein), were highly enriched during selection.

A cDNA enriched by selection with TNF- α was found to encode full-length Gadd45 β , a member of the Gadd45 family of inducible factors⁸ associated with cell-cycle control and DNA repair⁹. *gadd45 β* was strongly and rapidly induced by TNF- α in wild-type mouse embryo fibroblasts (MEFs), but not in RelA^{-/-} cells (Fig. 1a). This mirrored the expression of *ikb α* , a target of NF- κ B¹. *gadd45 β* was also upregulated by TNF- α in parental and Neo 3DO T cells, but not in 3DO clones expressing I κ B α M, a variant of I κ B α that blocks activation of NF- κ B¹⁰ (Fig. 1b; see also Supplementary Information Fig. 1). *gadd45 β* was also induced by NF- κ B after treatment with daunorubicin or phorbol 12-myristate-13-acetate (PMA) plus ionomycin (Fig. 1d and c, respectively).

In both MEFs and 3DO cells, expression of the other family members *gadd45 α /gadd45* and *gadd45 γ /oig37/cr6/grp17* (ref. 11) was independent of NF- κ B (Fig. 1b–d; and data not shown). Thus, unlike these latter genes, *gadd45 β* is a TNF- α -inducible gene and a physiological target of NF- κ B. Mutational analyses confirmed the presence of three functional κ B elements in the *gadd45 β* promoter (R.J., E.D.S. and G.F., manuscript in preparation).

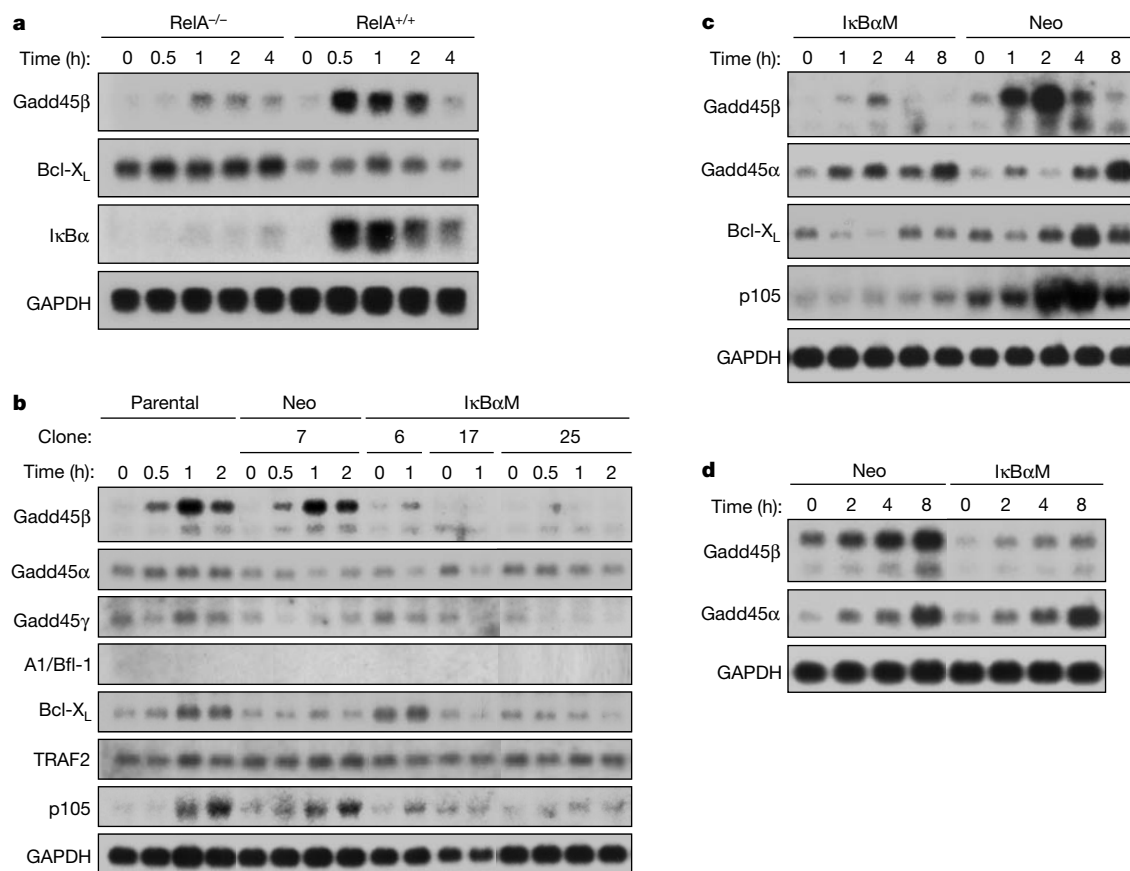


Figure 1 *gadd45* is a physiological target of NF- κ B. **a**, Northern blots with RNA from untreated and TNF- α -treated (1,000 U ml⁻¹) RelA^{-/-} and RelA^{+/+} MEFs. **b–d**, Northern blots with RNA from 3DO-IkB α M and 3DO-Neo clones treated with TNF- α (1,000 U ml⁻¹), PMA (50 ng ml⁻¹) plus ionomycin (1 μ M), or daunorubicin (0.5 μ M). The probes are indicated.

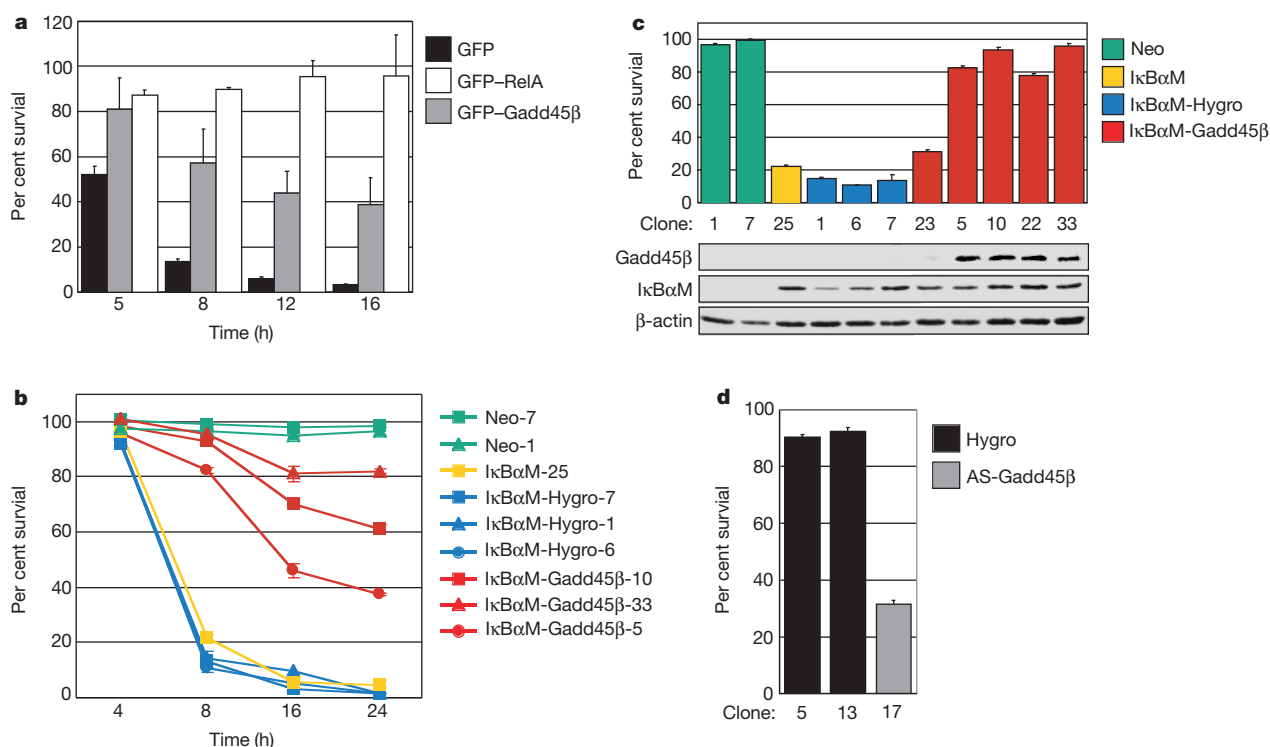


Figure 2 Gadd45 β antagonizes TNF-R-induced apoptosis in NF- κ B null cells. **a**, Survival of transfected RelA^{-/-} cells after treatment with TNF- α plus CHX. **b**, Survival of Neo, IkB α M, IkB α M-Hygro and IkB α M-Gadd45 β 3DO clones after TNF- α treatment. **c**, Eight-hour time point of **b** with two additional IkB α M-Gadd45 β clones (top), and western blots showing IkB α M and Gadd45 β protein levels (bottom). **d**, Survival of AS-Gadd45 β and Hygro 3DO clones after a 14-h treatment with TNF- α (1,000 U ml⁻¹) plus CHX (0.1 μ g ml⁻¹). In **a–d**, values represent the mean of three independent experiments.

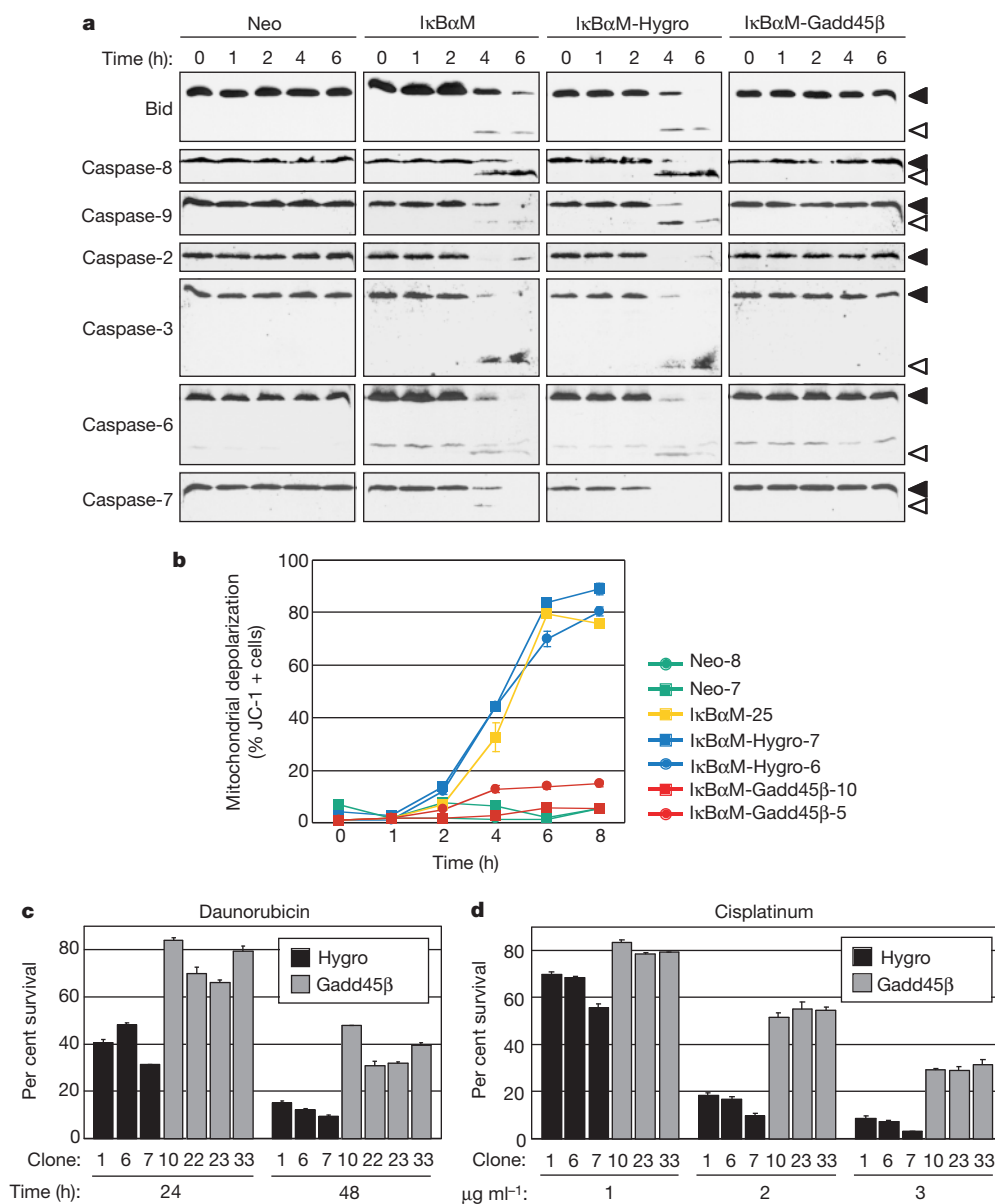


Figure 3 Gadd45 β blocks apoptotic pathways in NF- κ B null cells. **a**, Western blots showing processing of Bid and procaspases in 3DO clones after treatment with TNF- α (50 U ml $^{-1}$). Filled and open arrowheads indicate unprocessed and processed proteins, respectively. **b**, Mitochondrial depolarization in 3DO clones treated as in **a**. Values express

the percentage of JC-1 $^{+}$ cells in each culture. **c**, **d**, Survival of IkB α M-Gadd45 β and IkB α M-Hygro clones after treatment with 0.025 μ M daunorubicin (Sigma) or cisplatin (VHAplus) for 24 h. In **b**–**d**, values represent the mean of three independent experiments.

On expression in RelA $^{-/-}$ MEFs, a fusion protein of green fluorescent protein (GFP), GFP-Gadd45 β , exhibited marked anti-apoptotic activity against TNF- α , whereas GFP protein did not (Fig. 2a). Cell rescue was virtually complete with GFP-RelA. To determine whether Gadd45 β activity extended to other cell types, we examined NF- κ B-deficient (IkB α M) 3DO clones. These clones were highly sensitive to TNF- α -induced killing, and toxicity correlated with levels of IkB α M protein (see Supplementary Information Fig. 2). Expression of Gadd45 β markedly rescued 3DO-IkB α M-25 clones from PCD by TNF- α , with the rates of survival correlating with the presence of high protein levels of Gadd45 β (Fig. 2b and c; see also Supplementary Information Fig. 3).

Each of eight additional IkB α M-Hygro clones that were tested remained susceptible to TNF- α -induced killing, whereas in 27 other IkB α M-Gadd45 β clones resistance to apoptosis correlated with levels of Gadd45 β (data not shown). At early time points, protection by Gadd45 β was complete (Fig. 2b, see 8-h time point), and in

some clones was nearly complete even at 24 h, indicating that Gadd45 β can temporarily substitute for NF- κ B complexes. Notably, IkB α M-Gadd45 β and IkB α M-Hygro clones expressed comparable levels of IkB α M proteins (Fig. 2c), which were sufficient to block NF- κ B activation by TNF- α (Supplementary Information Fig. 1). Thus, as also seen in RelA $^{-/-}$ cells, Gadd45 β blocks apoptotic pathways by acting downstream of NF- κ B.

In NF- κ B-deficient 3DO cells, TNF- α activated caspase-8, -9, -2, -6, -3 and -7, as assessed by western blots for caspases and the caspase-8 substrate Bid 7 , and by caspase-specific proteolytic assays (Fig. 3a and Supplementary Information Fig. 4). As previously reported, these events were blocked completely by NF- κ B 12 (Neo). Caspase activation was also suppressed in IkB α M-Gadd45 β -10 cells, which expressed high levels of Gadd45 β . In these cells, as well as in NF- κ B-competent cells (Neo), mitochondrial depolarization—a key event in apoptosis 7 —was also blocked completely (Fig. 3b). Thus, Gadd45 β abrogates TNF- α -induced caspase activa-

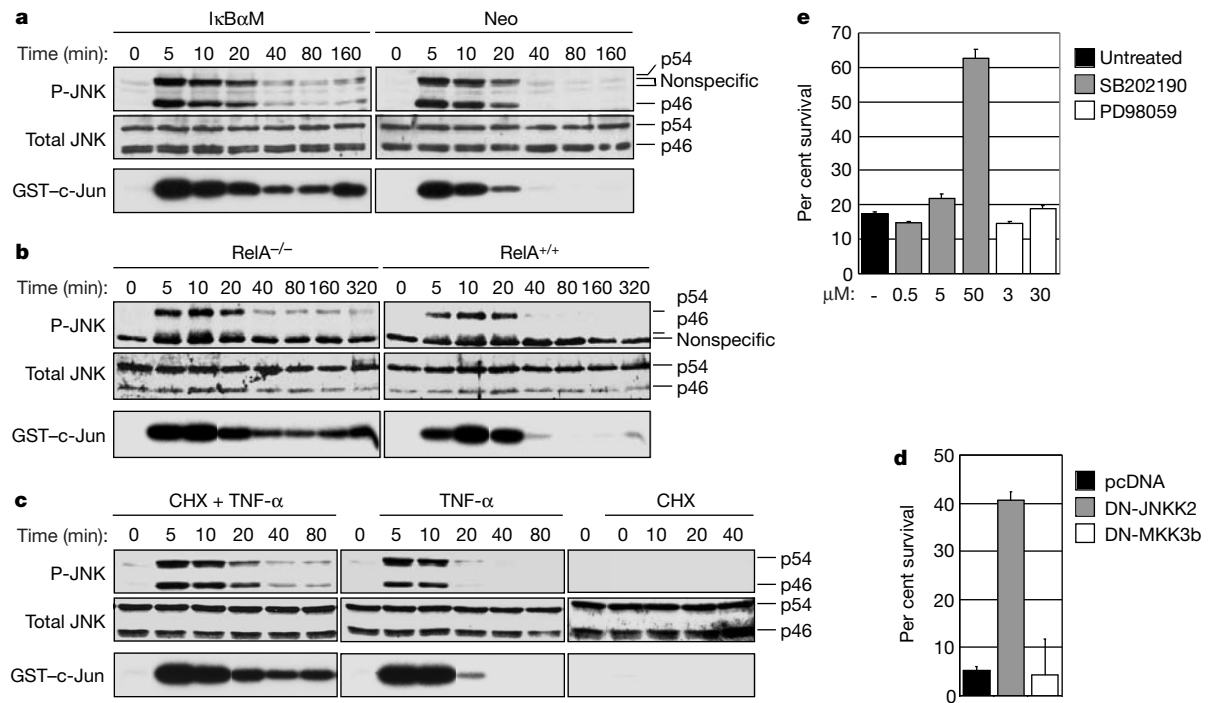


Figure 4 The inhibition of JNK represents a protective mechanism by NF-κB. **a**, **b**, JNK activation by TNF-α in 3DO-IκBαM and 3DO-Neo clones, and RelA^{-/-} and RelA^{+/+} MEFs, respectively. **c**, JNK activation in parental 3DO cells treated with TNF-α, TNF-α plus CHX, or CHX alone. **a–c**, Western blots with antibodies against phosphorylated (P) or total JNK (top and middle) and JNK kinase assays (bottom). **d**, Survival of transfected RelA^{-/-} MEFs

after a 10-h treatment with TNF-α plus CHX. DN, dominant negative. pcDNA (Invitrogen), empty vector control. **e**, Survival of 3DO-IκBαM cells after treatment with TNF-α or TNF-α plus MAPK inhibitors. In **d**, **e**, values represent the mean of three independent experiments.

tion and mitochondrial depolarization in NF-κB null cells, thereby recapitulating the effects of the transcription factor on death pathways. Gadd45β also protected cells against PCD induced by the genotoxic agents daunorubicin and cisplatin (Fig. 3c and d), which can trigger apoptotic pathways independently of TNF-Rs⁷.

To verify that the cytoprotection mediated by Gadd45β was not due to its overexpression, we generated 3DO clones expressing *gadd45β* in an antisense orientation. Despite their ability to activate NF-κB, cells expressing high levels of antisense *gadd45β* (AS-Gadd45β) exhibited a marked susceptibility to TNF-α-induced killing when compared with control cells (Fig. 2d). Cytotoxicity correlated with the levels of antisense messenger RNA (data not shown). Thus, Gadd45β is required to antagonize TNF-R-induced apoptosis, and cytoprotection is a physiological function of Gadd45β.

With regard to TNF-R, the NF-κB protective function has been associated with the induction of Bcl-2 family members Bcl-X_L and A1/Bfl-1, and the simultaneous upregulation of TRAF (TNF receptor-associated factor) 1/2 and c-IAP (cellular inhibitor of apoptosis protein) 1/2 (refs 12–14). However, TRAF2, c-IAP1 and Bcl-X_L are not induced significantly by TNF-α in various cell types and are found at near-normal levels in several NF-κB null cells^{12,15,16}. Moreover, these Bcl-2 family members or TRAFs and c-IAPs can only partly inhibit PCD in NF-κB-deficient cells^{12,13,17}. In addition, expression of TRAF1 and A1/Bfl-1 is restricted to certain tissues^{13,18}, and many cell types express TRAF1 in the absence of TRAF2 (ref. 18)—a factor that is needed to recruit TRAF1 to TNF-R⁷.

Other putative NF-κB targets, including A20 and IEX-1L, either do not protect NF-κB null cells against TNF-α or have been questioned to have anti-apoptotic activity^{6,19}. We found that in both MEFs and 3DO cells Bcl-X_L and TRAF2 were only marginally induced by TNF-α, whereas expression of A1/Bfl-1 mRNA was undetectable (Fig. 1a and b; data not shown). In RelA^{-/-} MEFs, basal levels of Bcl-X_L were even higher than in controls (Fig. 1a), and in 3DO cells expression did not correlate with resistance to TNF-α

(Fig. 1b; see also Supplementary Information Fig. 2). Thus, these genes cannot be general effectors of the protective function of NF-κB, and in these systems Gadd45β seems to be more relevant to the NF-κB control of apoptosis.

TNF-R also activates the extracellular signal-regulated kinases (ERK)-1/2, the JNK1/2/3 and the p38 protein kinases (p38α, -β, -γ, -δ)²⁰. It has been shown that apoptosis signal-regulating kinase 1 (ASK1) null cells, which cannot sustain JNK and p38 activation after TNF-R triggering, exhibit a profound defect in the apoptotic response to the cytokine²¹. This prompted us to analyse whether the activation of JNK and/or p38 is altered in NF-κB null cells. Whereas in Neo clones TNF-α-induced JNK signalling returned to near-basal levels by 40 min, in 3DO-IκBαM cells JNK signalling remained sustained throughout the time course (Fig. 4a, western blots and kinase assays; see also Fig. 5a). Qualitatively similar findings were obtained with two additional IκBαM and five Neo clones (data not shown), and with RelA^{-/-} and RelA^{+/+} MEFs (Fig. 4b). Conversely, ERK and p38 were unaffected (data not shown). Thus, NF-κB is required to terminate JNK signalling that is triggered by TNF-R.

As also shown by others²², cycloheximide (CHX) treatment impaired the downregulation of TNF-α-induced JNK activation (Fig. 4c), indicating that this process requires factors that are newly induced and/or rapidly turned over. Therefore, we tested whether Gadd45β mediated the inhibitory effect of NF-κB on JNK. In IκBαM-Gadd45β cells, JNK activation by TNF-α was markedly impaired (Fig. 5a; see also Fig. 5c); Gadd45β also suppressed the activation of JNK by sorbitol and hydrogen peroxide (Fig. 5b). The blockade was specific, as Gadd45β did not affect activation of either ERK or p38 (Fig. 5c). Consistent with these findings, antisense inhibition of endogenous Gadd45β led to a prolonged activation of JNK after triggering by TNF-R (Fig. 5d), showing that this factor is required for JNK inhibition. Together, these results indicate that the downregulation of JNK by NF-κB involves the induction of

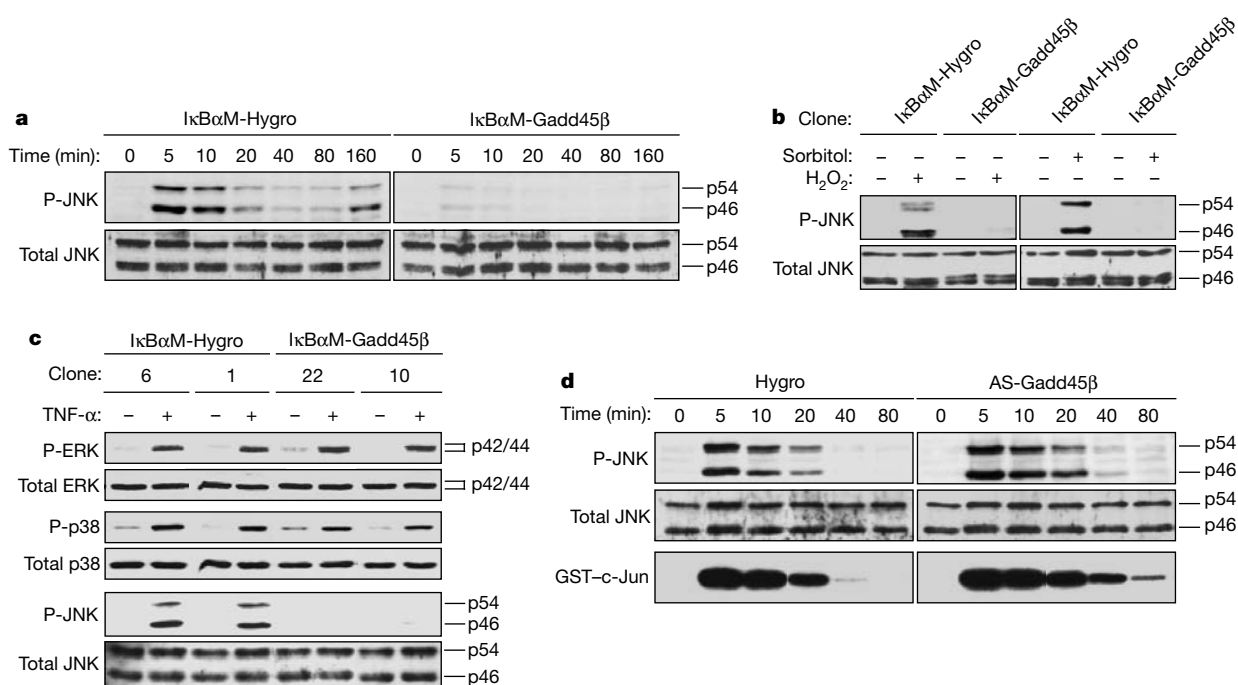


Figure 5 Gadd45β is a physiological inhibitor of JNK signalling. **a**, Western blots showing JNK activation by TNF-α in IkBαM-Hygro and IkBαM-Gadd45β 3DO clones. Similar results were obtained with four additional IkBαM-Gadd45β and three IkBαM-Hygro clones (**c**; and data not shown). **b**, JNK activation by hydrogen peroxide (H₂O₂, 600 μM)

and sorbitol (0.3 M) in IkBαM-Hygro and IkBαM-Gadd45β clones. Treatments were for 30 min. **c**, Western blots showing ERK, p38 and JNK phosphorylation in 3DO clones treated with TNF-α for 5 min. **d**, Western blots (top and middle) and kinase assays (bottom) showing JNK activation in AS-Gadd45β and Hygro clones treated with TNF-α.

Gadd45β, and that the inhibition of JNK may represent a mechanism for the protective activity of Gadd45β.

Thus, persistent activation of JNK correlates with cytotoxicity, whereas suppression of JNK correlates with cytoprotection. Indeed, we found that inhibition of JNK by dominant-negative JNKK2/MKK7 (ref. 23) rescued RelA null cells from TNF-α-induced killing, whereas inhibition of p38 by dominant-negative MKK3b²⁴ had no impact on cell survival (Fig. 4d). Similar findings were obtained in 3DO-IkBαM cells, in which inhibition of JNK was achieved by using well-characterized pharmacological blockers: PD98059 and low concentrations of SB202190, which specifically block ERK and p38, respectively, did not antagonize TNF-α cytotoxicity, whereas high concentrations of SB202190, which blocks both p38 and JNK²⁵, markedly enhanced cell survival (Fig. 4e). Together, the data indicate that JNK is crucial to PCD induced by TNF-R, and that one mechanism by which NF-κB protects cells is the downregulation of the JNK cascade by means of Gadd45β.

Here we have identified Gadd45β as a physiological regulator of the apoptotic response to TNF-α. *gadd45β* is upregulated rapidly by the cytokine through a mechanism that requires NF-κB, is essential to antagonize TNF-α-induced killing, and blocks PCD in NF-κB null cells. We also found that the NF-κB anti-apoptotic function depends on the suppression of JNK. This cross-talk between the NF-κB and JNK pathways involves the transcriptional activation of *gadd45β*. These findings represent a mechanism for cell protection mediated by NF-κB and establish a pro-apoptotic role for the persistent activation of JNK in the response to TNF-α. Consistent with our results, apoptosis has been associated previously with continuous activation of JNK²⁶. Inhibition of JNK might also explain the protective activity of Gadd45β against genotoxic-stress-induced apoptosis, which requires JNK³. Because expression of *gadd45β* is induced by chemotherapeutic drugs (Fig. 1d), this gene might also contribute to NF-κB-dependent cancer chemoresistance, and thereby represent a potential therapeutic target².

The precise steps regulated by JNK during apoptosis and the mechanisms by which Gadd45β affects JNK activation remain to be

determined. As Gadd45β has not been reported to have enzymatic activity (for instance, phosphatase activity), this factor might interfere with protein kinase function, either directly or indirectly. Of note, Gadd45α is known to inhibit Cdc2/cyclin B1 kinase by binding to the core enzyme and displacing the cyclin B1 coactivator⁹. Thus, it is possible that Gadd45β inhibits the JNK cascade by disrupting analogous cofactor-kinase interactions or by interfering with direct associations between JNK signalling modules²⁰.

Notably, Gadd45 proteins were shown to associate with MEKK4/MTK1 and proposed to function as initiators of JNK and p38 signalling⁸. Two subsequent studies have shown, however, that expression of Gadd45 proteins does not induce JNK or p38 in various cell lines and that endogenous products make no contribution to the activation of these kinases by stress^{27,28}. Our findings are consistent with these latter studies and, furthermore, establish a role for Gadd45β as an inhibitor of JNK. TNF-R-triggered JNK activation precedes the induction of Gadd45β (Figs 1b and 4a) and is enhanced and prolonged by the ablation of NF-κB or the presence of CHX (Fig. 4a–c)—two conditions that block the induction of Gadd45β (see Fig. 1a and b). Foremost, the expression of Gadd45β effectively inhibits JNK activation by several stimuli (Fig. 5a–c), whereas antisense *gadd45β* enhances it (Fig. 5d). Nevertheless, the ability of Gadd45 proteins to bind to MEKK4 does support the existence of a link between these proteins and kinases in the JNK pathway. Another study has also implicated Gadd45γ in the activation of JNK, as well as p38, in T-helper cells²⁹. Further studies on the mechanisms by which Gadd45 proteins interfere with JNK signalling will be necessary to resolve these issues. □

Methods

Library enrichment

We prepared cDNA from TNF-α-treated NIH-3T3 cells and inserted it directionally into the pLTP vector³. For the library enrichment, about 4.5 × 10⁶ library clones were transfected into RelA^{-/-} cells by the spheroplast fusion method, and cells were treated with TNF-α (100 U ml⁻¹) and CHX (0.25 mg ml⁻¹) for 21 h. Episomal DNA was extracted from

adherent cells, amplified and retransfected. Four consecutive cycles of library selection were completed.

Cytotoxic treatments, kinase assays and antibodies

Plasmids were transiently transfected into RelA^{-/-} cells, and after 24 h cultures were treated with CHX (0.1 µg ml⁻¹) with or without TNF-α (1,000 U ml⁻¹; Peprotech). In Fig. 4d, plasmids were co-transfected with pEGFP (Clontech). At the indicated times, we counted and analysed adherent cells by flow cytometry (FCM) to determine the numbers of live GFP-positive cells. With 3DO clones, apoptosis was assessed by staining with propidium iodide (Sigma) followed by FCM analysis, as described³⁰. Unless otherwise stated, 3DO cells were treated with 25 U ml⁻¹ TNF-α. In Fig. 4e, 3DO-IκBαM cells were pretreated with mitogen-activated protein kinase (MAPK) inhibitors (Calbiochem) for 30 min and then incubated with either TNF-α (25 U ml⁻¹) or PBS for an additional 12 h. In both RelA^{-/-} MEFs and 3DO cells, percentages of live cells in experimental cultures were calculated relative to the number of live cells present in mock-treated cultures. We performed kinase assays with recombinant glutathione S-transferase (GST)-c-Jun and anti-JNK antibodies (Pharmingen), as described²³. In all cases, cells were treated with 1,000 U ml⁻¹ TNF-α. In Fig. 4c, CHX treatments (10 µg ml⁻¹) were carried out for 30 min in addition to the indicated time. We used the following antibodies: phospho-JNK (Fig. 4b), IκBαM and β-actin from Santa Cruz Biotechnology; Bid, caspase-6, -7 and -9, and total and phospho-JNK, -ERK and -p38 from Cell Signalling Technology; caspase-8 from Alexis; caspase-2 and -3 from R&D Systems. The anti-Gadd45β antibody was generated against an N-terminal peptide of murine Gadd45β.

Caspase activity and transmembrane potential

We performed *in vitro* fluorimetric assays of caspase activity as described³⁰, by incubating cell lysates with amino trifluoromethyl coumarin (ATC)-labelled caspase-specific peptides (Bachem): zVDVAD (caspase-2), zDEVD (caspases-3/7), zVEID (caspase-6), zETD (caspase-8) and Ac-LEHD (caspase-9). The mitochondrial transmembrane potential was measured by using the fluorescent dye JC-1 (Molecular Probes) and FCM, according to the manufacturer's instructions.

Received 29 June; accepted 24 August 2001.

- Ghosh, S., May, M. J. & Kopp, E. B. NF-κB and Rel proteins: evolutionarily conserved mediators of immune responses. *Annu. Rev. Immunol.* **16**, 225–260 (1998).
- Baldwin, A. S. Control of oncogenesis and cancer therapy resistance by the transcription factor NF-κB. *J. Clin. Invest.* **107**, 241–246 (2001).
- Davis, R. J. Signal transduction by the JNK group of MAP kinases. *Cell* **103**, 239–252 (2000).
- Vairapandi, M., Balliet, A. G., Fornace J. J. Jr, Hoffman, B. & Liebermann, D. A. The differentiation primary response gene MyD118, related to GADD45, encodes for a nuclear protein which interacts with PCNA and p21^{WAF1/CIP1}. *Oncogene* **12**, 2579–2594 (1996).
- Vito, P., Lacana, E. & D'Adamo, L. Interfering with apoptosis: Ca²⁺-binding protein ALG-2 and Alzheimer's disease gene ALG-3. *Science* **271**, 521–525 (1996).
- Beg, A. A. & Baltimore, D. An essential role for NF-κB in preventing TNF-α-induced cell death. *Science* **274**, 782–787 (1996).
- Budihardjo, I., Oliver, H., Lutter, M., Luo, X. & Wang, X. Biochemical pathways of caspase activation during apoptosis. *Annu. Rev. Cell Dev. Biol.* **15**, 269–290 (1999).
- Takekawa, M. & Saito, H. A family of stress-inducible GADD45-like proteins mediate activation of the stress-responsive MTK1/MEKK4 MAPKKK. *Cell* **95**, 521–530 (1998).
- Zhan, Q. *et al.* Association with Cdc2 and inhibition of Cdc2/Cyclin B1 kinase activity by the p53-regulated protein Gadd45. *Oncogene* **18**, 2892–2900 (1999).
- Van Antwerp, D. J., Martin, S. J., Kafri, T., Green, D. R. & Verma, I. M. Suppression of TNF-α-induced apoptosis by NF-κB. *Science* **274**, 787–789 (1996).
- Zhang, W. *et al.* CR6: A third member in the MyD118 and Gadd45 gene family which functions in negative growth control. *Oncogene* **18**, 4899–4907 (1999).
- Wang, C.-Y., Mayo, M. W., Korneluk, R. G., Goeddel, D. V. & Baldwin, A. S. Jr NF-κB antiapoptosis: Induction of TRAF1 and TRAF2 and c-IAP1 and c-IAP2 to suppress caspase-8 activation. *Science* **281**, 1680–1683 (1998).
- Zong, W. X., Edelstein, L. C., Chen, C., Bash, J. & Gelinas, C. The prosurvival Bcl-2 homolog Bfl-1/A1 is a direct transcriptional target of NF-κB that blocks TNF-α-induced apoptosis. *Genes Dev.* **13**, 382–387 (1999).
- Chen, C., Edelstein, L. C. & Gelinas, C. The Rel/NF-κB family directly activates expression of the apoptosis inhibitor Bcl-x_L. *Mol. Cell. Biol.* **20**, 2687–2695 (2000).
- Schwenzel, R. *et al.* The human tumor necrosis factor (TNF-) receptor-associated factor 1 gene (TRAF1) is up-regulated by cytokines of the TNF-ligand family and modulates TNF-induced activation of NF-κB and c-Jun N-terminal kinase. *J. Biol. Chem.* **274**, 19368–19374 (1999).
- Chu, Z.-L. *et al.* Suppression of tumor necrosis factor-induced cell death by inhibitor of apoptosis c-IAP2 is under NF-κB control. *Proc. Natl Acad. Sci. USA* **94**, 10057–10062 (1997).
- Duriez, P. J., Wong, F., Dorovini-Zis, K., Shahidi, R. & Karsan, A. A1 functions at the mitochondria to delay endothelial apoptosis in response to tumor necrosis factor. *J. Biol. Chem.* **275**, 18099–18107 (2000).
- Zapata, J. M. *et al.* TNF-R-associated factor family protein expression in normal tissues and lymphoid malignancies. *J. Immunol.* **165**, 5084–5096 (2000).
- Arlt, A. Expression of the NF-κB target gene IEX-1 (p22/PRG1) does not prevent cell death but instead triggers apoptosis in HeLa cells. *Oncogene* **20**, 69–76 (2001).
- Chang, L. & Karin, M. Mammalian MAP kinase signalling cascades. *Nature* **410**, 37–40 (2001).
- Tobiume, K. *et al.* ASK1 is required for sustained activations of JNK-p38 MAP kinases and apoptosis. *EMBO Rep.* **2**, 222–228 (2001).
- Guo, Y.-L., Baysal, K., Kang, B., Yang, L.-J. & Williamson, J. R. Correlation between sustained c-Jun N-terminal protein kinase activation and apoptosis induced by tumor necrosis factor-α in rat mesangial cells. *J. Biol. Chem.* **273**, 4027–4034 (1997).
- Lu, X., Nemoto, S. & Lin, A. Identification of c-Jun NH2-terminal protein kinase (JNK)-activating kinase 2 as an activator of JNK but not p38. *J. Biol. Chem.* **272**, 24751–24754 (1997).

- Huang, S. *et al.* Apoptosis signaling pathway in T cells is composed of ICE/Ced-3 family proteases and MAP kinase kinase 6. *Immunity* **6**, 739–749 (1997).
- Jacinto, E., Werlen, G. & Karin, M. Cooperation between Syk and Rac1 leads to synergistic JNK activation in T lymphocytes. *Immunity* **8**, 31–41 (1998).
- Chen, Y. R. & Tan, T. H. The c-jun N-terminal kinase pathway and apoptotic signaling. *Int. J. Oncol.* **16**, 651–662 (2000).
- Shaulian, E. & Karin, M. Stress-induced JNK activation is independent of Gadd45 induction. *J. Biol. Chem.* **274**, 29595–29598 (1999).
- Wang, X., Gorospe, M. & Holbrook, N. J. Gadd45 is not required for activation of c-Jun N-terminal kinase or p38 during acute stress. *J. Biol. Chem.* **274**, 29599–29602 (1999).
- Lu, B. *et al.* Gadd45γ mediates the activation of the p38 and JNK pathways and cytokine production in effector T_H1 cells. *Immunity* **14**, 583–590 (2001).
- Stegh, A. H. *et al.* Identification of the cytolinker plectin as a major early *in vivo* substrate for caspase 8 during CD95- and tumor necrosis factor receptor-mediated apoptosis. *Mol. Cell. Biol.* **15**, 5665–5679 (2000).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

Acknowledgements

We thank M. Peter, A. Ma, H. Singh, C.-R. Wang, P. Ashton-Rickardt, M. Rosner and C. G. Pham for comments on the manuscript. We are grateful to L. D'Adamo for advice on the manuscript and the 'death trap' system, and for pLTP. We thank D. Baltimore and A. Hoffmann for the RelA^{-/-} MEFs; A. Lin for dominant-negative JNK2; J. Han for MKK3b; D. Liebermann and A. Balliet for Gadd45β reagents; and S. Rua for help with manuscript preparation. Supported in part by the Damon Runyon Scholar Award of the Cancer Research Fund and an NIH grant.

Correspondence and requests for materials should be addressed to G.F. (e-mail: gfranzos@midway.uchicago.edu).

Inhibition of JNK activation through NF-κB target genes

Guilin Tang[†], Yuzuru Minemoto[†], Benjamin Dibling^{*}, Nicole H. Purcell^{*}, Zhiwei Li[‡], Michael Karin[‡] & Anning Lin^{*}

^{*} Ben May Institute for Cancer Research, Committee on Cancer Biology, University of Chicago, 5841 S. Maryland Avenue, MC 6027, Chicago, Illinois 60637, USA

[‡] Laboratory of Gene Regulation and Signal Transduction, Department of Pharmacology, University of California, San Diego, 9500 Gilman Drive, La Jolla, California 92093-0636, USA

[†] These authors contributed equally to the work

The proinflammatory cytokine tumour necrosis factor-α (TNF-α) regulates immune responses, inflammation and programmed cell death (apoptosis)^{1–4}. The ultimate fate of a cell exposed to TNF-α is determined by signal integration between its different effectors, including IκB kinase (IKK), c-Jun N-terminal protein kinase (JNK) and caspases¹. Activation of caspases is required for apoptotic cell death⁵, whereas IKK activation inhibits apoptosis through the transcription factor NF-κB, whose target genes include caspase inhibitors^{1,6–10}. JNK activates the transcription factor c-Jun/AP-1, as well as other targets^{11–16}. However, the role of JNK activation in apoptosis induced by TNF-α is less clear^{17,18}. It is unknown whether any crosstalk occurs between IKK and JNK, and, if so, how it affects TNF-α-induced apoptosis. We investigated this using murine embryonic fibroblasts that are deficient in either the IKKβ catalytic subunit of the IKK complex or the RelA/p65 subunit of NF-κB. Here we show that in addition to inhibiting caspases, the IKK/NF-κB pathway negatively modulates TNF-α-mediated JNK activation, partly through NF-κB-induced X-chromosome-linked inhibitor of apoptosis (XIAP)^{7,9}. This negative crosstalk, which is specific to TNF-α signalling and does not affect JNK activation by interleukin-1 (IL-1), contributes to inhibition of apoptosis.

IKKβ is required for activation of NF-κB by TNF-α, whereas

IKK α is dispensable^{6,19,20}. To determine whether interfering with the IKK/NF- κ B signalling pathway affects JNK activation by TNF- α , we used wild-type (WT), IKK α -deficient (*Ikk α ^{-/-}*)¹⁹ and IKK β -deficient (*Ikk β ^{-/-}*)²⁰ mouse embryonic fibroblasts (Fig. 1a). After treatment with TNF- α , the activities of IKK and JNK were measured by immune-complex kinase assays^{15,21}. Activation of IKK by TNF- α was transient and robust in WT fibroblasts and *Ikk α ^{-/-}* cells, but severely impaired in *Ikk β ^{-/-}* cells (Fig. 1b). By contrast, JNK activation was transient in WT and *Ikk α ^{-/-}* fibroblasts but long lasting in *Ikk β ^{-/-}* cells (Fig. 1c). This sustained activation did not result from increased expression of either p46^{JNK} (Fig. 1c) or p54^{JNK} (data not shown). Transfection of *Ikk β ^{-/-}* cells with an IKK β expression vector restored transient JNK activation in response to TNF- α (Fig. 1d). Loss of IKK β did not affect TNF- α -mediated activation of p38 MAPK (Fig. 1e) or ERK (Fig. 1f).

To investigate the role of NF- κ B in IKK-mediated JNK inhibition, we used RelA-deficient (*RelA^{-/-}*) mouse embryonic fibroblasts (Fig. 2a) lacking the major activating subunit of NF- κ B²². Immune-complex kinase assays showed that TNF- α induced sustained JNK activation in *RelA^{-/-}* fibroblasts (Fig. 2b). Even basal JNK activity was slightly higher in *RelA^{-/-}* cells than in WT controls (Fig. 2b). The sustained activation was not the result of increased JNK expression (Fig. 2b). Transfection of *RelA^{-/-}* fibroblasts with a RelA expression vector restored transient JNK activation in response to TNF- α stimulation (Fig. 2c).

Because JNK is a key regulator of c-Jun activity¹¹, its negative regulation by the IKK/NF- κ B pathway might affect c-Jun-mediated transcription. Indeed, TNF- α stimulated the transcriptional activity of a GAL4-c-Jun chimera, in which the GAL4 DNA binding domain was fused to the c-Jun transactivation domain¹⁵, by 20-

fold in *RelA^{-/-}* cells but only 7-fold in WT controls (Fig. 2d). Transfection of *RelA^{-/-}* fibroblasts with a RelA expression vector, which activated an NF- κ B reporter gene (Fig. 2d), abrogated the stimulation of GAL4-c-Jun activity by TNF- α (Fig. 2d). This suggests that the augmented c-Jun activation is due to the loss of NF- κ B-mediated JNK inhibition.

To determine whether gene expression induced by NF- κ B is required to block JNK activation, WT fibroblasts were treated with TNF- α and the protein synthesis inhibitor cycloheximide (CHX) or CHX alone. In the presence of CHX, TNF- α induced prolonged JNK activation (Fig. 2e). The prolonged activation of JNK by TNF- α in CHX-treated WT fibroblasts was less sustained than in *RelA^{-/-}* cells (Fig. 2b). This was probably due to incomplete inhibition of *de novo* protein synthesis by CHX at the concentration (10 μ g ml⁻¹) used (data not shown). The sustained activation of JNK by TNF- α plus CHX did not result from increased expression of JNK, as examined by immunoblotting (data not shown). Although CHX can activate JNK in some cell types, it had a marginal effect on JNK activity in mouse fibroblasts (Fig. 2e). It seems that NF- κ B-mediated suppression of JNK activation requires *de novo* synthesis of JNK inhibitors.

Dephosphorylation is a probable mechanism for inactivation of JNK^{23,24}. We examined whether NF- κ B inhibits JNK activation through induction of JNK phosphatases. Incubation of purified active JNK with extracts from either non-treated WT or *RelA^{-/-}* cells resulted in partial JNK inactivation (Fig. 3a, ~50% reduction). Treatment of either WT or *RelA^{-/-}* cells with TNF- α did not increase JNK phosphatase activity (Fig. 3a). These results, however, do not exclude the possibility that NF- κ B may induce phosphatases that inactivate upstream JNK activators. We therefore examined JNK

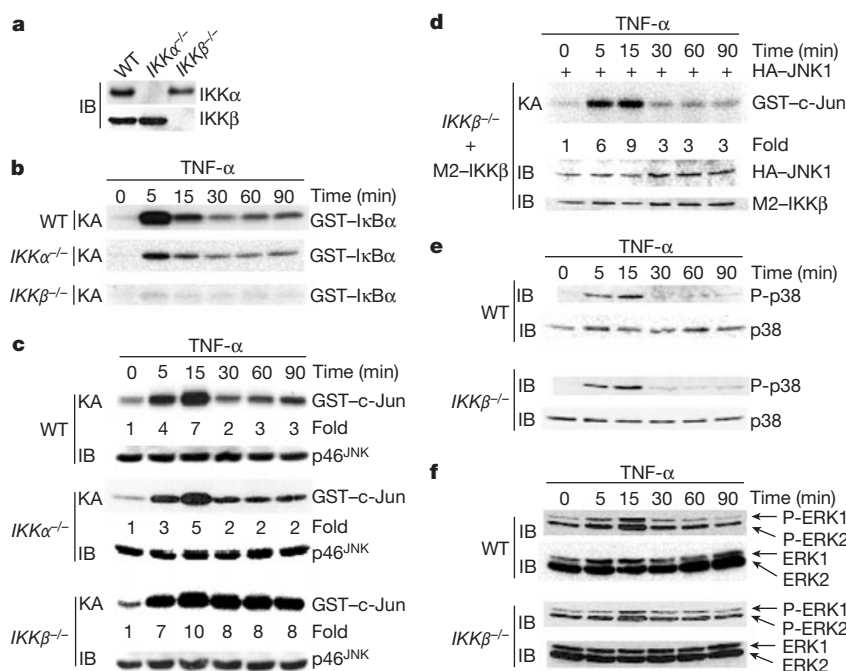


Figure 1 Genetic disruption of IKK β results in sustained JNK activation by TNF- α . **a**, Immunoblot (IB) analysis of IKK α and IKK β expression in WT, *Ikk α ^{-/-}* and *Ikk β ^{-/-}* fibroblasts with anti-IKK α (Santa Cruz) and anti-IKK β (Upstate) antibodies, respectively. **b**, Activation of IKK by TNF- α is impaired in *Ikk β ^{-/-}* fibroblasts. WT, *Ikk α ^{-/-}* and *Ikk β ^{-/-}* fibroblasts were treated with TNF- α (20 ng ml⁻¹) as indicated. IKK was immunoprecipitated with anti-IKK γ antibody (Santa Cruz) and its activity measured by immune-complex kinase assay (KA) using glutathione *S*-transferase (GST)-I κ B α (1–54) as substrate²¹. **c**, Activation of JNK by TNF- α is prolonged in *Ikk β ^{-/-}* fibroblasts. The same cell lysates as in **b** were examined for JNK activation by immune-complex kinase assay using GST-c-Jun(1–79) as substrate¹⁵. Substrate phosphorylation was quantified by

phosphorimager, and fold stimulation is indicated. JNK content was analysed by immunoblotting with anti-JNK antibody (Pharmingen). **d**, Transfection of *Ikk β ^{-/-}* fibroblasts with an IKK β expression vector restores transient JNK activation by TNF- α . *Ikk β ^{-/-}* cells were cotransfected with expression vectors encoding Flag-tagged(M2)-IKK β (2 μ g) and HA-JNK1 (1 μ g). After 36 h, cells were treated with TNF- α (20 ng ml⁻¹) as indicated. The activity of HA-JNK1 and expression of HA-JNK1 and M2-IKK β was determined. **e**, **f**, Normal activation of p38 or ERK by TNF- α in *Ikk β ^{-/-}* fibroblasts. The same cell lysates as in **b** were analysed by immunoblotting for activation of p38 (**e**) or ERK (**f**) using corresponding anti-phospho antibodies or pan antibodies (New England Biolabs).

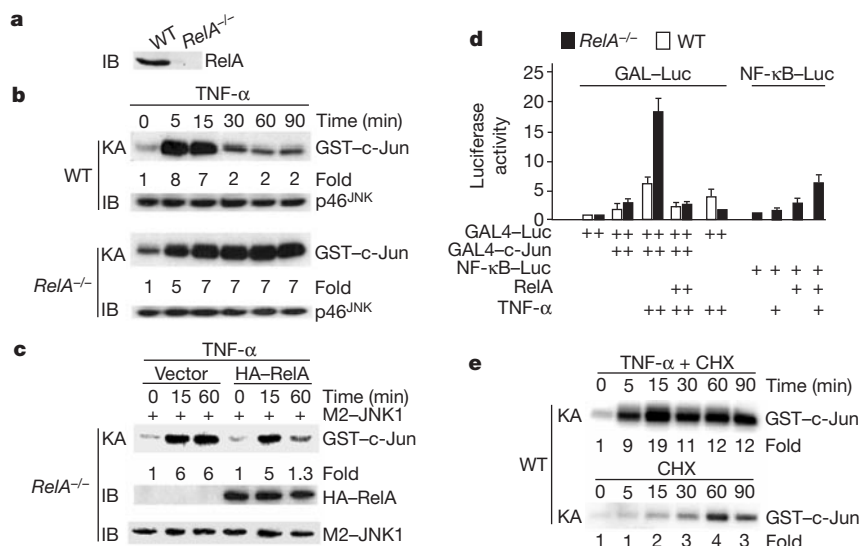


Figure 2 NF-κB mediates the inhibitory effect of IKK on JNK. **a**, Immunoblot analysis of RelA expression in WT and *RelA*^{-/-} fibroblasts with anti-RelA antibody (Zymed). **b**, Activation of JNK by TNF-α is sustained in *RelA*^{-/-} fibroblasts. WT and *RelA*^{-/-} fibroblasts were treated with TNF-α (20 ng ml⁻¹) as indicated, and collected. JNK activity and expression were examined. **c**, Transfection of *RelA*^{-/-} fibroblasts with a RelA expression vector restores transient JNK activation by TNF-α. *RelA*^{-/-} cells were cotransfected with expression vectors encoding M2-JNK1 (1 μg) and either HA-RelA (1 μg) or empty vector (1 μg). After 36 h, cells were treated with TNF-α (20 ng ml⁻¹), as indicated. The activity and expression of M2-JNK1 and expression of HA-RelA were

determined. **d**, Expression of RelA inhibits c-Jun transcriptional activity. WT and *RelA*^{-/-} fibroblasts were cotransfected with a 5× GAL4-Luc reporter plasmid (1 μg), a 2× NF-κB-Luc reporter plasmid (0.25 μg) and expression vectors encoding GAL4-c-Jun(1-223) (2 ng) and RelA (5 ng), as indicated. After 20 h, cells were treated with TNF-α (20 ng ml⁻¹) for 10 h. Cells were collected and relative luciferase activity was determined^{15,21}. The results are presented as means ± standard errors and represent four independent experiments done in duplicate. **e**, TNF-α induces prolonged JNK activation in the presence of CHX. WT fibroblasts were treated with TNF-α (20 ng ml⁻¹) and CHX (10 μg ml⁻¹) or CHX alone. JNK activity was examined.

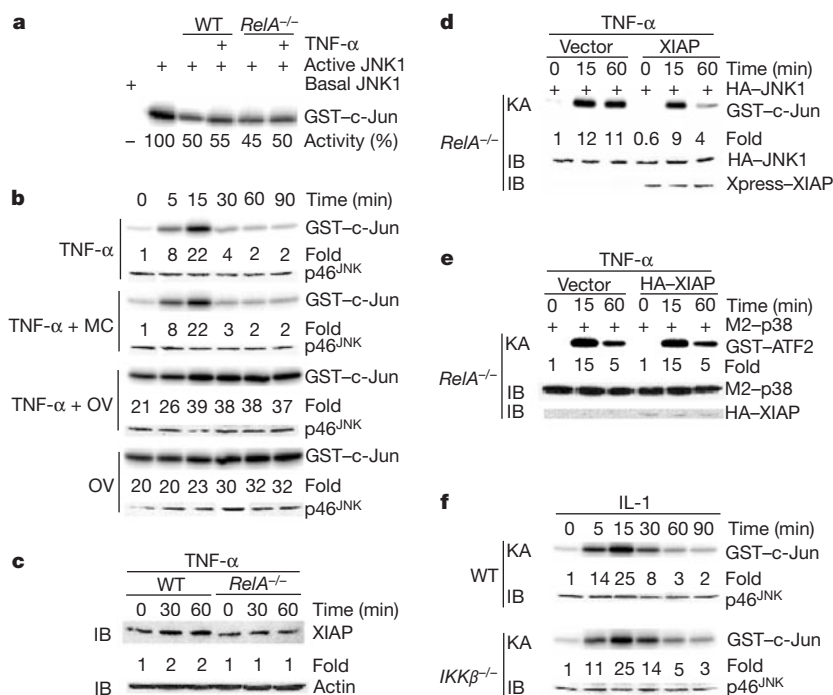


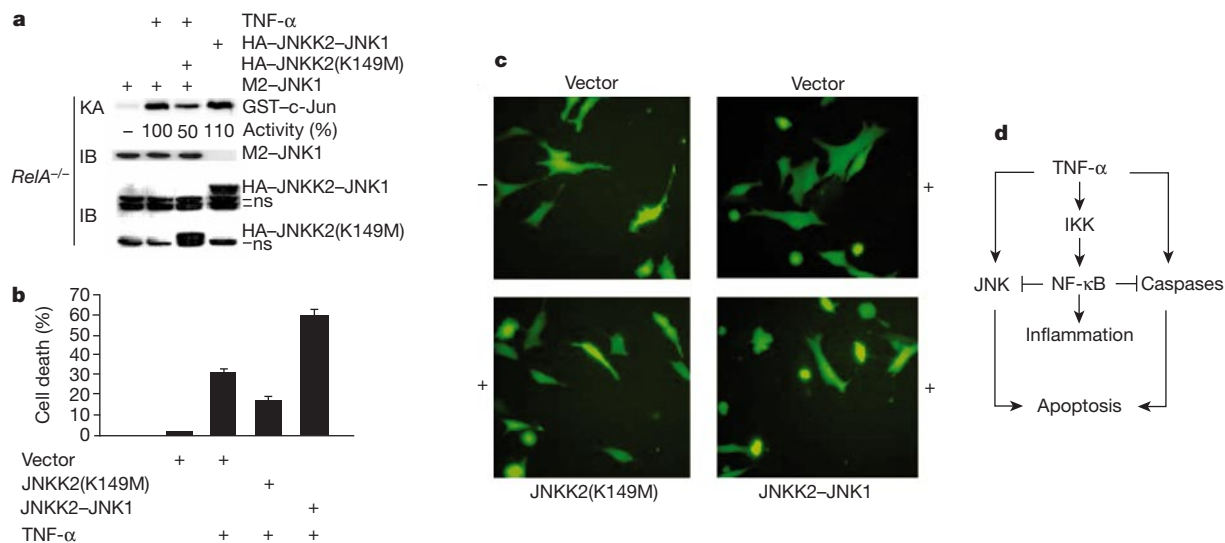
Figure 3 Inhibition of JNK activation by NF-κB is specific to TNF-α signalling and involves XIAP, but not phosphatases. **a**, JNK isolated by immunoprecipitation from non-stimulated (basal) or ultraviolet-irradiated (active) HeLa cells was incubated with untreated or TNF-α-treated WT or *RelA*^{-/-} fibroblast extracts for 1 h in phosphatase buffer²⁴. After extensive washing, JNK activity was examined. The activity of active JNK incubated with buffer alone was given an arbitrary value of 100%. **b**, WT fibroblasts were treated with or without microcystin (MC, 200 nM) or vanadate (OV, 500 μM) for 1 h before treatment with or without TNF-α (20 ng ml⁻¹). JNK activity and expression were examined. JNK activity was expressed relative to the JNK activity of untreated cells, which was given an arbitrary value of 1. **c**, WT and *RelA*^{-/-} cells were treated with TNF-α (20 ng ml⁻¹) as indicated and

expression of XIAP was analysed by immunoblotting using anti-XIAP antibody (PharMingen). The same blot was analysed for expression of actin. **d**, *RelA*^{-/-} cells were cotransfected with expression vectors for HA-JNK1 (2 μg) and either Xpress-tagged XIAP (Xpress-XIAP) (0.5 μg) or empty vector (0.5 μg). After 40 h, cells were treated with or without TNF-α (20 ng ml⁻¹) and the activity and expression of HA-JNK1 and Xpress-XIAP were determined. **e**, *RelA*^{-/-} cells were transiently transfected with M2-p38 (2 μg) and either HA-XIAP (0.5 μg) or empty vector (0.5 μg). After 40 h, cells were treated with or without TNF-α (20 ng ml⁻¹) and the activity and expression of M2-p38 and HA-XIAP were determined. **f**, WT and *IKKβ*^{-/-} fibroblasts were treated with IL-1 (2 ng ml⁻¹) as indicated. JNK activity and expression was determined as described in Fig. 1c.

These results led us to search for other candidates that could function as NF- κ B-induced inhibitors of JNK activation. We tested several known NF- κ B target proteins, including XIAP, c-IAP1, c-FLIP_L, Bcl-X_L and Bcl-2, and found that only transient expression of XIAP inhibited JNK activation by TNF- α in HeLa cells (see Supplementary Information Fig. 2). Furthermore, TNF- α induced XIAP expression in WT, but not *RelA*^{-/-} fibroblasts (Fig. 3c), confirming that the XIAP gene is targeted by NF- κ B⁹. Transfection of *RelA*^{-/-} fibroblasts with a XIAP expression vector restored transient JNK activation in response to TNF- α and modestly decreased TNF- α -stimulated JNK activity (Fig. 3d). These results are consistent with the observation that in *RelA*^{-/-} cells, TNF- α -induced JNK activation was only modestly enhanced whereas the JNK inactivation that occurs after 30 min of TNF- α treatment was profoundly blocked (Figs 1c and 2b). This suggests that XIAP may attenuate TNF- α -mediated JNK activation. Although it has been reported that overexpression of XIAP activates JNK through an unknown mechanism²⁶, we found that overexpression of XIAP simply enhanced expression levels of cotransfected JNK, resulting in higher basal JNK activity. Expression of small amounts of XIAP,

Sustained JNK activation is suggested to contribute to apoptotic cell death^{23,28}. The prevention of sustained JNK activation by NF- κ B may contribute to its anti-apoptotic activity. To examine this possibility, we used the dominant negative JNKK2(K149M) mutant²⁹ and the constitutively active JNKK2–JNK1 fusion protein³⁰. Expression of JNKK2(K149M) significantly attenuated JNK activation by TNF- α , whereas the JNKK2–JNK1 fusion protein had kinase activity similar to that of TNF- α -activated JNK (Fig. 4a). TNF- α induced apoptosis in *RelA*^{-/-} fibroblasts (Fig. 4b and c; 30% cell death). Expression of JNKK2(K149M) suppressed TNF- α -induced cell death (Fig. 4b and c, 18% cell death), as did expression of the dominant negative JNKK1(257/261AA) mutant (16% cell death) (data not shown). Conversely, TNF- α -induced apoptosis was doubled in *RelA*^{-/-} fibroblasts expressing the JNKK2–JNK1 fusion protein (Fig. 4b and c, 60% cell death). The augmentation of TNF- α -induced cell killing was probably due to the sustained JNK activity of the JNKK2–JNK1 fusion protein, because expression of a kinase-defective JNKK2(K149M)–JNK1 fusion protein had no effect (data not shown). The JNKK2–JNK1 fusion protein alone had no detectable apoptotic effect in *RelA*^{-/-} fibroblasts (data not shown).

Our results show that IKK negatively modulates JNK activity, most probably through the induction of NF- κ B target genes encoding proteins such as XIAP, which interfere with TNF- α .



HA-JNK2(K149M) mutant or empty vector (2.8 μg each). Cells were treated with or without TNF- α (20 ng ml $^{-1}$) for 6 h, stained with Hoechst dye and visualized by fluorescence microscopy. The death rate of transfected (GFP positive) cells was calculated by counting 12 randomly selected areas (total 40–50 cells per area). Results are presented as means $\pm$ standard errors and represent two individual experiments. The differences in the cell death were statistically significant ($P < 0.05$, Student's t -test). +, TNF- α treated, –, untreated. **d**, Schematic illustration of negative regulation of JNK by IKK. See text for details.

mediated, but not IL-1-mediated, JNK activation. This highly specific form of crosstalk allows NF- κ B activation to suppress TNF- α -induced apoptosis through inhibition of both caspases and sustained JNK activation (Fig. 4d). Because apoptosis is an anti-inflammatory process¹ and sustained JNK activation promotes apoptosis, suppression of sustained JNK activation by NF- κ B might enhance TNF- α -induced inflammation (Fig. 4d). □

Methods

Generation of immortalized mouse fibroblasts

WT, *Ikk α ^{-/-}* and *Ikk β ^{-/-}* mouse fibroblasts have been described previously^{19,20}. Cells were immortalized by the 3T3 protocol.

In vitro phosphatase assays

WT or *RelA^{-/-}* cells were treated with or without TNF- α (20 ng ml⁻¹, 60 min). Preparation of cell extracts and phosphatase assays were described previously²⁴.

Apoptosis assays

RelA^{-/-} fibroblasts were cotransfected with expression vectors encoding green fluorescent protein (GFP) and either haemagglutinin-tagged JNK2–JNK1 (HA–JNK2–JNK1), HA–JNK2(K148M) or empty vector at a ratio of 1:4. Under these conditions, cells expressing GFP also expressed either HA–JNK2–JNK1 or HA–JNK2(K148M). Cells were treated with or without TNF- α for 6 h and apoptosis was monitored by Hoechst staining for nuclear condensation.

Received 19 June; accepted 21 September 2001.

- Baud, V. & Karin, M. Signal transduction by tumor necrosis factor and its relatives. *Trends Cell Biol.* **11**, 372–377 (2001).
- Tartaglia, L. A. & Goeddel, D. V. Two TNF receptors. *Immunol. Today* **13**, 151–153 (1992).
- Tracey, K. J. & Cerami, A. Tumor necrosis factor: an updated review of its biology. *Crit. Care Med.* **21**, S415–S422 (1993).
- Locksley, R. M., Killeen, N. & Lenardo, M. J. The TNF and TNF receptor superfamilies: integrating mammalian biology. *Cell* **104**, 487–501 (2001).
- Thornberry, N. A. & Lazebnik, Y. Caspases: enemies within. *Science* **281**, 1312–1316 (1998).
- Rothwarf, D. M. & Karin, M. The NF- κ B activation pathway: a paradigm in information transfer from membrane to nucleus. *Science STKE* [online] (cited 26 October 1999) (http://www.stke.org/cgi/content/full/OC_sigtrans;1999/5/rel1) (1999).
- Duckett, C. *et al.* A conserved family of cellular genes related to the baculovirus *iap* gene and encoding apoptosis inhibitors. *EMBO J.* **15**, 2685–2694 (1996).
- Chu, Z. L. *et al.* Suppression of tumor necrosis factor-induced cell death by inhibitor of apoptosis c-IAP2 is under NF- κ B control. *Proc. Natl Acad. Sci. USA* **94**, 10057–10062 (1997).
- Stehlik, C. *et al.* Nuclear factor (NF)- κ B-regulated X-chromosome-linked *iap* gene expression protects endothelial cells from tumor necrosis factor alpha-induced apoptosis. *J. Exp. Med.* **188**, 211–216 (1998).
- Wang, C. Y., Mayo, M. W., Korneluk, R. G., Goeddel, D. V. & Baldwin, A. S. Jr NF- κ B antiapoptosis: induction of TRAF1 and TRAF2 and c-IAP1 and c-IAP2 to suppress caspase-8 activation. *Science* **281**, 1680–1683 (1998).
- Chang, L. & Karin, M. Mammalian MAP kinase signalling cascades. *Nature* **410**, 37–40 (2001).
- Hibi, M., Lin, A., Smeal, T. & Karin, M. Identification of an oncoprotein- and UV-responsive protein kinase that binds and potentiates the c-Jun activation domain. *Genes Dev.* **7**, 2135–2148 (1993).
- Derijard, B. *et al.* JNK1: a protein kinase stimulated by UV light and Ha-Ras that binds and phosphorylates the c-Jun activation domain. *Cell* **76**, 1025–1037 (1994).
- Kyriakis, J. M. *et al.* The stress-activated protein kinase subfamily of c-Jun kinases. *Nature* **369**, 156–160 (1994).
- Lin, A. *et al.* Identification of a dual specificity kinase that activates the Jun kinases and p38-Mpk2. *Science* **268**, 286–290 (1995).
- Minden, A. *et al.* Differential activation of ERK and JNK mitogen-activated protein kinases by Raf-1 and MEKK. *Science* **266**, 1719–1723 (1994).
- Liu, Z. G., Hsu, H., Goeddel, D. V. & Karin, M. Dissection of TNF receptor 1 effector functions: JNK activation is not linked to apoptosis while NF- κ B activation prevents cell death. *Cell* **87**, 565–576 (1996).
- Natoli, G. *et al.* Activation of SAPK/JNK by TNF receptor 1 through a noncytotoxic TRAF2-dependent pathway. *Science* **275**, 200–203 (1997).
- Hu, Y. *et al.* Abnormal morphogenesis but intact IKK activation in mice lacking the IKK α subunit of I κ B kinase. *Science* **284**, 316–320 (1998).
- Li, Z. W. *et al.* The IKK β subunit of I κ B kinase (IKK) is essential for nuclear factor κ B activation and prevention of apoptosis. *J. Exp. Med.* **189**, 1839–1845 (1999).
- Nemoto, S., DiDonato, J. A. & Lin, A. Coordinate regulation of I κ B kinases by mitogen-activated protein kinase kinase kinase 1 and NF- κ B-inducing kinase. *Mol. Cell. Biol.* **18**, 7336–7343 (1998).
- Beg, A. A., Sha, W. C., Bronson, R. T., Ghosh, S. & Baltimore, D. Embryonic lethality and liver degeneration in mice lacking the RelA component of NF- κ B. *Nature* **376**, 167–170 (1995).
- Guo, Y.-L., Kang, B. & Williamson, J. R. Correlation between sustained c-Jun N-terminal protein kinase activation and apoptosis induced by tumor necrosis factor- α in rat mesangial cells. *J. Biol. Chem.* **273**, 4027–4034 (1998).
- Cavigelli, M., Li, W. W., Lin, A., Su, B., Yoshioka, K. & Karin, M. The tumor promoter arsenites stimulates AP-1 activity by inhibiting a JNK phosphatase. *EMBO J.* **15**, 6269–6279 (1996).
- Javelaud, D. & Besancon, F. NF- κ B activation results in rapid inactivation of JNK in TNF- α -treated Ewing sarcoma cells: a mechanism for the anti-apoptotic effect of NF- κ B. *Oncogene* **20**, 4365–4372 (2001).

- Sanna, M. G., Duckett, C. S., Richter, B. W., Thompson, C. B. & Ulevitch, R. J. Selective activation of JNK1 is necessary for the anti-apoptotic activity of hILP. *Proc. Natl Acad. Sci. USA* **95**, 6015–6020 (1998).
- Baud, V. *et al.* Signaling by proinflammatory cytokines: oligomerization of TRAF2 and TRAF6 is sufficient for JNK and IKK activation and target gene induction via an amino-terminal effector domain. *Genes Dev.* **13**, 1297–1308 (1999).
- Chen, Y.-R., Wang, X., Templeton, D., Davis, R. & Tan, T.-H. The role of c-Jun N-terminal kinase (JNK) in apoptosis induced by ultraviolet C and gamma radiation. Duration of JNK activation may determine cell death and proliferation. *J. Biol. Chem.* **270**, 31929–31936 (1996).
- Lu, X., Nemoto, S. & Lin, A. Identification of c-Jun NH2-terminal protein kinase (JNK)-activating kinase 2 as an activator of JNK but not p38. *J. Biol. Chem.* **272**, 24751–24754 (1997).
- Zheng, C., Xiang, J., Hunter, T. & Lin, A. The JNK2–JNK1 fusion protein acts as a constitutively active c-Jun kinase that stimulates c-Jun transcription activity. *J. Biol. Chem.* **274**, 28966–28971 (1999).

Supplementary Information accompanies the paper on Nature's website (<http://www.nature.com>).

Acknowledgements

We are grateful to D. Baltimore for providing us with *RelA^{-/-}* fibroblasts and C. S. Duckett for the XIAP construct; M. R. Rosner for critical reading of this manuscript and suggestions; and S. Nemoto for her initial finding when she was a member of Lin laboratory. This work was supported by grants from the National Institutes of Health (NIH), the American Cancer Society (ACS) and Department of Defense to A.L. Generation of IKK-deficient cells was supported by grants from the NIH and State of California Cancer Research Program to M.K., who is the Frank and Else Schilling-ACS Research Professor.

Correspondence and requests for materials should be addressed to A.L. (email: alin@huggins.bsd.uchicago.edu).

The 7SK small nuclear RNA inhibits the CDK9/cyclin T1 kinase to control transcription

Zhiyuan Yang*, Qingwei Zhu*†, Kunxin Luo*† & Qiang Zhou*

* Department of Molecular and Cell Biology, University of California at Berkeley; and † Life Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA

The human positive transcription elongation factor P-TEFb, consisting of a CDK9/cyclin T1 heterodimer, functions as both a general and an HIV-1 Tat-specific transcription factor^{1,2}. P-TEFb activates transcription by phosphorylating RNA polymerase (Pol) II, leading to the formation of processive elongation complexes. As a Tat cofactor, P-TEFb stimulates HIV-1 transcription by interacting with Tat and the transactivating responsive (TAR) RNA structure located at the 5' end of the nascent viral transcript³. Here we identified 7SK, an abundant and evolutionarily conserved small nuclear RNA (snRNA) of unknown function^{4,5}, as a specific P-TEFb-associated factor. 7SK inhibits general and HIV-1 Tat-specific transcriptional activities of P-TEFb *in vivo* and *in vitro* by inhibiting the kinase activity of CDK9 and preventing recruitment of P-TEFb to the HIV-1 promoter. 7SK is efficiently dissociated from P-TEFb by treatment of cells with ultraviolet irradiation and actinomycin D. As these two agents have been shown to significantly enhance HIV-1 transcription and phosphorylation of Pol II (refs 6–8), our data provide a mechanistic explanation for their stimulatory effects. The 7SK/P-TEFb interaction may serve as a principal control point for the induction of cellular and HIV-1 viral gene expression during stress-related responses. Our studies demonstrate the involvement of an snRNA in controlling the activity of a Cdk–cyclin kinase.

To identify nuclear factors that can interact with and regulate the activity of P-TEFb, Flag-tagged CDK9 and its associated factors

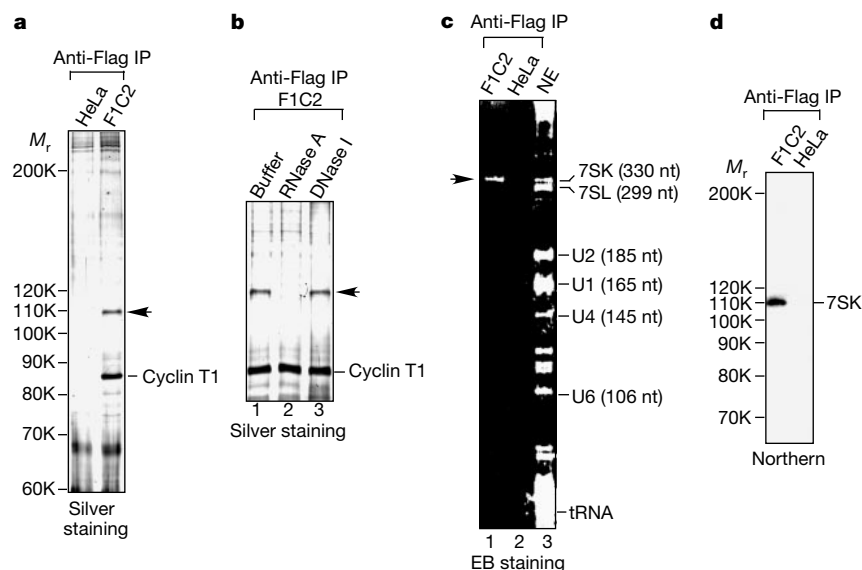


Figure 1 CDK9 associates with the 7SK snRNA. **a**, CDK9-Flag and its associated factors (anti-Flag immunoprecipitates (IP)) were affinity purified from the nuclear extract (NE) of F1C2 cells, and analysed by SDS–polyacrylamide gel electrophoresis and silver staining. HeLa nuclear extract was used in a parallel procedure for control. The arrow indicates an unknown band that has a relative molecular mass of 110,000 (M_r 110K). **b**, Affinity-purified CDK9-Flag and its associated factors were incubated with RNase A, DNase I, or buffer alone, and analysed by silver staining. **c**, RNA was extracted from HeLa

nuclear extract (lane 3) or the anti-Flag immunoprecipitates derived from F1C2 and HeLa nuclear extract (lanes 1 and 2), and analysed on a denaturing polyacrylamide gel stained with ethidium bromide (EB). nt, nucleotides. Arrows in **b** and **c** are the same as in **a**. **d**, RNA recovered from the anti-Flag immunoprecipitates derived from F1C2 and HeLa nuclear extract was analysed by northern hybridization with full-length 7SK antisense RNA as a probe.

were affinity purified from the nuclear extract of an engineered HeLa cell line (F1C2 cells) that stably expressed CDK9-Flag⁹. Analysis of the purified material by silver staining detected cyclin T1 and a novel band with a relative molecular mass of 110,000 (M_r 110K) derived from F1C2 but not the parental HeLa cells (Fig. 1a). Coomassie blue could not stain the 110K band, and the yellowish, silver-stained colour was different from that of the brown colour typical of proteins, indicating that the band may not be a protein. Indeed, treatment of the affinity-purified CDK9-Flag preparation with RNase A, but not DNase I, eliminated the 110K band (Fig. 1b), suggesting that it may contain a CDK9-associated RNA molecule.

RNA was extracted from the CDK9-Flag preparation and analysed on a denaturing gel with small RNAs recovered from HeLa nuclear extract as markers (Fig. 1c). The identities of some of these HeLa RNAs were pre-determined by oligonucleotide-directed RNase H digestion¹⁰. The CDK9-Flag preparation contained a single RNA species that co-migrated with the 7SK RNA, comprising 330 nucleotides, in HeLa extract. On transcription by RNA Pol III, the mammalian 7SK RNA is an abundant (approximately 2×10^5 per cell) and evolutionarily conserved snRNA of unknown function^{4,11}. Using full-length 7SK antisense RNA as a probe, northern hybridization was performed to confirm that the CDK9-associated 110K RNA was 7SK (Fig. 1d). Sequencing of the complementary DNA copy of this RNA revealed a complete match with the published 7SK sequences¹².

To investigate the role of 7SK in transcription, it was quantitatively removed from HeLa nuclear extract (Fig. 2a) using an immobilized 2'-O-methyl (2'-OMe) RNA oligonucleotide complementary to an exposed region in 7SK (residues 221–241 (ref. 4)). Notably, the 7SK small nuclear ribonucleoprotein particle (snRNP) associated with the 2'-OMe RNA beads contained both cyclin T1 and CDK9 (Fig. 2b, lanes 1 and 2), indicating an association of 7SK with the CDK9/cyclin T1 heterodimer in HeLa nuclear extract. Approximately twice the amount of cyclin T1 and CDK9 was detected in the mock-depleted extract than in the extract that was

depleted of 7SK (lanes 3–6), suggesting that about 50% of P-TEFb may be stably associated with 7SK.

We next compared the abilities of the mock- and 2'-OMe RNA-depleted HeLa nuclear extracts to transcribe templates pSV40EP-G400 and pHIVΔTAR-G100 (ref. 13) in the same reaction. Removal of 7SK and its associated P-TEFb had no effect on transcription proximal to the promoter of the 400-nucleotide G-less cassette (G400), which was driven by the SV40 early promoter, or on transcription distal to the promoter of a G100 cassette, driven by the HIV-1 promoter (Fig. 2c); furthermore, there was no effect on Tat activation of HIV-1 transcription (data not shown). Thus, the P-TEFb bound to 7SK did not contribute to the transcriptional activity of HeLa nuclear extract.

If the P-TEFb bound to 7SK is inactive in transcription, we asked whether this might be due to 7SK having an inhibitory effect on the function of P-TEFb. To disrupt 7SK, oligonucleotide-directed RNase H digestion¹⁰ of 7SK was performed in the nuclear extract of F1C2 cells. Treatment with 221–241A—an antisense deoxyoligonucleotide targeted against 7SK—but not a control oligonucleotide, caused cleavage of full-length 7SK (330 nucleotides) into two fragments of approximately 220 and 90 nucleotides, respectively (Fig. 2d). The integrity of the targeted region (nucleotides 221–241) appeared to be critical for the binding of 7SK to P-TEFb, as very little of the cleaved 7SK fragments were associated with the affinity-purified CDK9-Flag/cyclin T1. Thus, treatment with 221–241A effectively created more P-TEFb that was not bound to 7SK (free P-TEFb), in the extract. Compared with both untreated and control oligonucleotide-treated extracts, the extract treated with 221–241A consistently yielded 2–3-fold more basal and Tat-activated HIV-1 transcription from templates pHIV+TAR-G400 (which contained the wild-type TAR element) and pHIVΔTAR-G100 (with a mutant TAR)¹³ (Fig. 2e). Given that only about 50% of P-TEFb associated with 7SK, the 2–3-fold increase in transcription was significant, and it suggested that 7SK might suppress the transcriptional activity of P-TEFb *in vitro*.

When 221–241A or a scrambled oligonucleotide, 221–241S, was

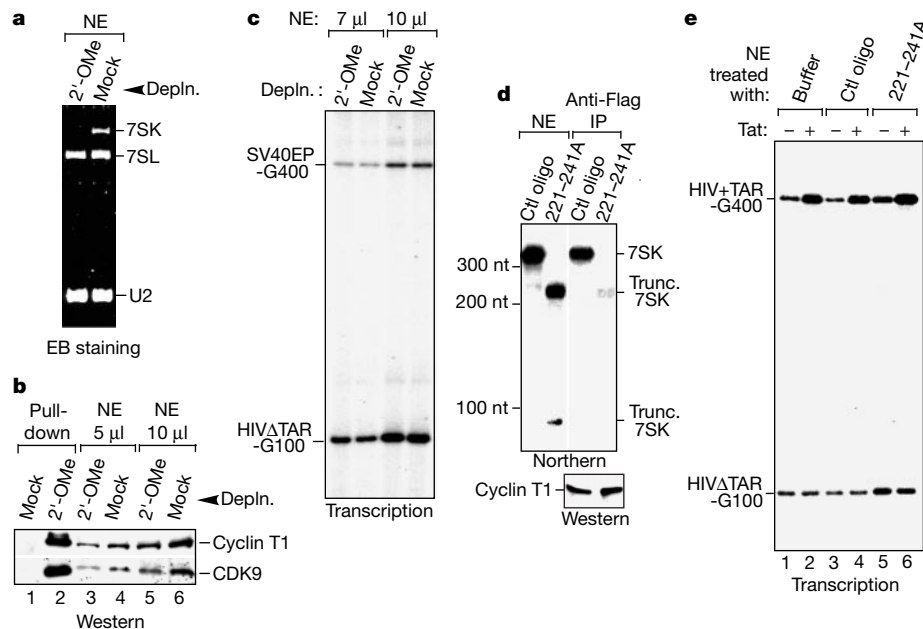


Figure 2 7SK binds to the CDK9/cyclin T1 heterodimer and inhibits P-TEFb transcriptional activity *in vitro*. **a**, HeLa nuclear extract (NE) was incubated with streptavidin agarose beads without (mock) or with immobilized 7SK antisense 2'-OMe RNA oligonucleotide. RNA recovered from the depleted extracts (depln.) was analysed on a gel stained with ethidium bromide (EB). **b**, Cyclin T1 and CDK9 associated with the mock or 2'-OMe RNA beads (lanes 1 and 2), or in the depleted extracts (lanes 3–6), were detected by immunoblotting. **c**, HeLa nuclear extract mock depleted or depleted of 7SK was incubated with templates pSV40EP-G400 and pHIVΔTAR-G100 in transcription reactions⁹. These reactions produced the indicated G-less transcripts.

d, FIC2 nuclear extract containing CDK9-Flag was incubated with the 7SK antisense deoxyoligonucleotide 221–241A or a control oligonucleotide (Ctl oligo). RNA recovered from the treated nuclear extract and its derivative anti-Flag immunoprecipitates was probed for 7SK by northern blotting. Cyclin T1 in the anti-Flag immunoprecipitates was detected by immunoblotting. Trunc., truncated 7SK. **e**, HeLa nuclear extract treated with the specified oligonucleotide was incubated with templates pHIV+TAR-G400 and pHIVΔTAR-G100 in transcription reactions with or without Tat. The G-less transcripts are indicated.

cotransfected with CDK9-Flag into HeLa cells, only 221–241A significantly reduced the binding of 7SK to CDK9-Flag (Fig. 3a), effectively creating more free P-TEFb in the cell. To determine whether 7SK also suppresses P-TEFb activity *in vivo*, the effect of 221–241A on the abilities of various promoters to transcribe a luciferase reporter gene was examined. Transfection of 221–241A, but not 221–241S, into HeLa cells increased transcription from all promoters tested, with the largest increase (roughly 9.5-fold) displayed by the HIV-1 long terminal repeat (LTR). Smaller increases were displayed by the SV40 early promoter, and the (Gal4)₅-thymidine kinase promoter, the transforming growth factor- β -responsive promoter p3TP (Fig. 3b). Similar results were also obtained in several other cell lines of diverse origins (data not shown). Finally, transfection of 221–241A into HeLa cells increased both basal and Tat-activated HIV-1 transcription (Fig. 3c). These experiments and the above *in vitro* transcriptional analyses reveal a general inhibitory effect of 7SK on P-TEFb transcriptional activity. Notably, HIV-1 LTR seems to be most sensitive to this inhibition, probably because it is regulated mainly at the stage of elongation and requires P-TEFb for both basal and Tat-activated transcription^{1,2}.

In addition to the region targeted by 221–241A, two other regions of 7SK (residues 11–31 and 95–114) are also potentially accessible to oligonucleotide-directed RNase H cleavage^{4,14}. Antisense and scrambled deoxyoligonucleotides specifically targeting these two regions were therefore transfected into HeLa cells together with an HIV-1 LTR luciferase construct. Compared with 221–241A, which increased significantly HIV transcription, a smaller increase was observed with the antisense oligonucleotide 95–114A, and no increase with oligonucleotide 11–31A was observed (Fig. 3d). The abilities of the three antisense oligonucleotides to increase transcription correlated exactly with their abilities to induce 7SK

cleavage (Fig. 3e, lanes 1–6) and to disrupt the interaction between 7SK and P-TEFb in the nuclear extract (lanes 7–12), providing further evidence of the inhibitory effect of 7SK on the transcriptional activity of P-TEFb.

7SK could inhibit the activity of P-TEFb by suppressing the kinase activity of CDK9/cyclin T1. To test this hypothesis, affinity-purified CDK9-Flag and its associated factors were divided into two halves, incubated respectively with RNase A and DNase I, and tested in kinase reactions containing purified RNA Pol II as a substrate⁹. RNase A degraded the CDK9-Flag-associated 7SK (Fig. 1b), and increased the kinase activity of CDK9-Flag by 3–4-fold, as seen by its increased autophosphorylation and phosphorylation of Pol II (Fig. 4a). This increase was significant given that only about 50% of the purified CDK9-Flag/cyclin T1 was associated with 7SK. In addition to RNase A, RNase H cleavage of 7SK directed by 221–241A also increased the kinase activity of CDK9 (data not shown). To analyse more specifically the activity of P-TEFb bound to 7SK, this complex was affinity purified from HeLa nuclear extract using the 7SK antisense 2'-OMe RNA beads (see above). Eluted with a displacement deoxyoligonucleotide¹⁵, the P-TEFb bound to 7SK was divided into three equal portions, incubated respectively with RNase A, DNase I or buffer alone, and analysed in kinase reactions. Once again, degradation of 7SK by RNase A significantly increased the kinase activity of CDK9 (Fig. 4b), revealing the inhibitory action of 7SK on CDK9/cyclin T1 kinase.

P-TEFb can be recruited to the pre-initiation complex (PIC) at the HIV-1 promoter and then travel with the elongating Pol II (refs 16, 17). The mechanism of recruitment is unclear although the interaction of cyclin T1 with the hypophosphorylated Pol II (ref. 18) could be responsible. To study the effect of 7SK on the association of P-TEFb with PIC, an immobilized HIV-1 promoter was incubated with FIC2 nuclear extract to isolate the promoter-bound

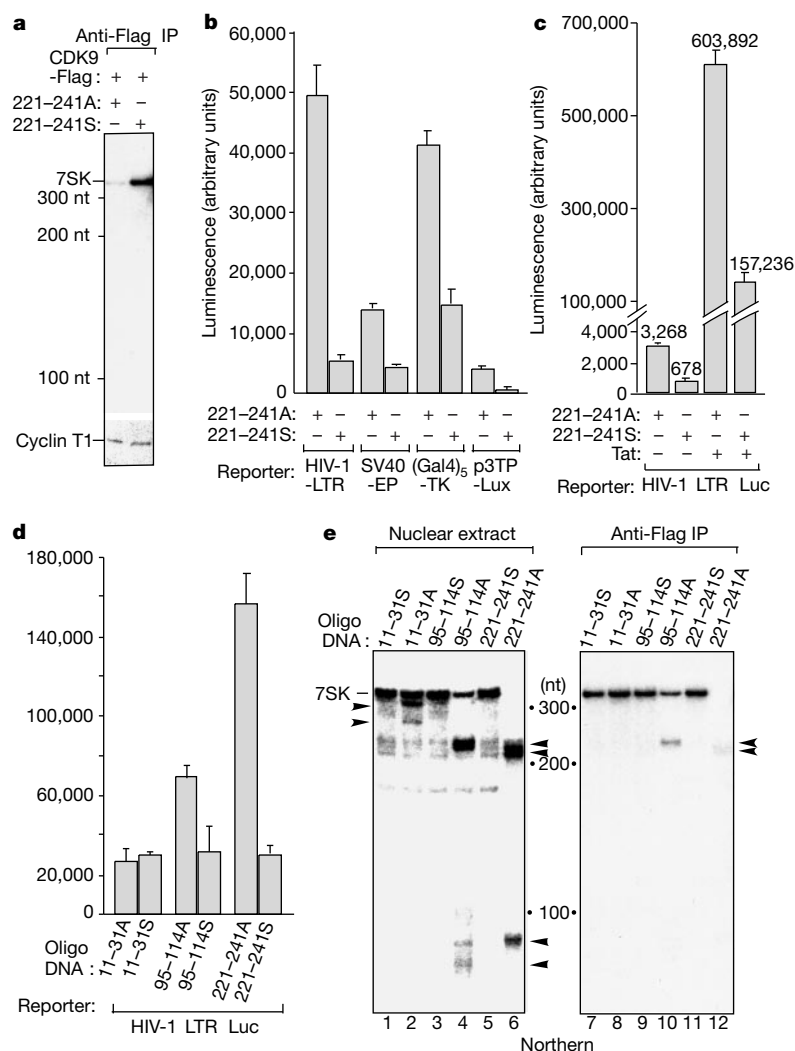


Figure 3 7SK inhibits P-TEFb transcriptional activity *in vivo*. **a**, HeLa cells were cotransfected with CDK9-Flag and the indicated deoxyoligonucleotide. 7SK and cyclin T1 associated with CDK9-Flag were affinity purified (anti-Flag IP) from the nuclear extract of the transfected cells and analysed by northern and immunoblotting, respectively. **b**, The indicated oligonucleotide and promoter-luciferase construct were cotransfected into HeLa cells and luciferase activities were measured 48 h later. EP, early promoter; TK, thymidine kinase; Lux, luciferase. **c**, The indicated oligonucleotide, an HIV-1 LTR luciferase

construct and a Tat-expressing construct were cotransfected into HeLa cells, and luciferase activities were measured 48 h later. **d**, HeLa cells were cotransfected and analysed as in **c**. The numbers associated with the various antisense (A) and scrambled (S) oligonucleotides denote the targeted regions of 7SK. **e**, Northern analysis of 7SK extracted from F1C2 nuclear extracts treated with the indicated oligonucleotides (lanes 1–6), and 7SK associated with CDK9-Flag affinity purified from the treated nuclear extracts (lanes 7–12). Arrows indicate truncated 7SK.

P-TEFb^{16,17}. Northern blotting was performed to compare the level of 7SK associated with the promoter-bound P-TEFb with that in the total P-TEFb affinity purified from the nuclear extract. When normalized by their cyclin T1 and CDK9 levels, the promoter-bound P-TEFb showed no 7SK, whereas abundant 7SK existed in the total P-TEFb preparation (Fig. 4c), suggesting that 7SK prevented the binding of P-TEFb to the HIV-1 promoter *in vitro*. To verify this *in vivo*, a chromatin immunoprecipitation (CHIP¹⁹) assay was performed to examine the interaction of P-TEFb with an integrated HIV-1 promoter in HeLa cells. Cells were cotransfected with CDK9-Flag and either 221–241A or 221–241S. As shown above, transfected 221–241A disrupted the 7SK/P-TEFb interaction and increased HIV-1 transcription. Notably, it also increased the association of CDK9-Flag with the HIV-1 promoter in these cells (Fig. 4d). Thus, the 7SK-P/TEFb interaction not only inhibited the kinase activity of P-TEFb, but also blocked the recruitment of P-TEFb to the HIV-1 promoter.

Certain agents that elicit SOS-like stress responses in mammalian cells can markedly enhance HIV-1 transcription in a manner

analogous to λ prophage induction in *Escherichia coli*^{6–8,20,21}. For instance, treatment of HeLa cells with ultraviolet irradiation or low levels of the global transcription inhibitor actinomycin D enhances HIV-1 transcription to levels similar to those obtained by Tat (refs 6, 8, and our own unpublished data). Notably, these stimulatory effects seemed to be caused by an enhanced phosphorylation of Pol II, probably by the CDK9/cyclin T1 kinase⁶. In light of these observations, we investigated the effect of ultraviolet irradiation and actinomycin D on the 7SK/P-TEFb interaction in nuclear extracts of the treated F1C2 cells. Consistent with their enhancement of HIV-1 transcription and Pol II phosphorylation, both agents significantly reduced the amount of 7SK associated with the affinity-purified CDK9-Flag (lanes 1 and 2 in Fig. 4e, f). Neither the level of total 7SK in the nuclear extract nor the CDK9/cyclin T1 interaction was affected (Fig. 4e, f). Actinomycin D is known to intercalate into duplex DNA and perhaps also RNA; however, direct incubation of this drug with nuclear extract did not disrupt the 7SK/P-TEFb interaction (data not shown), ruling out a direct effect. As for ultraviolet irradiation, 7SK was dissociated from P-TEFb as early as

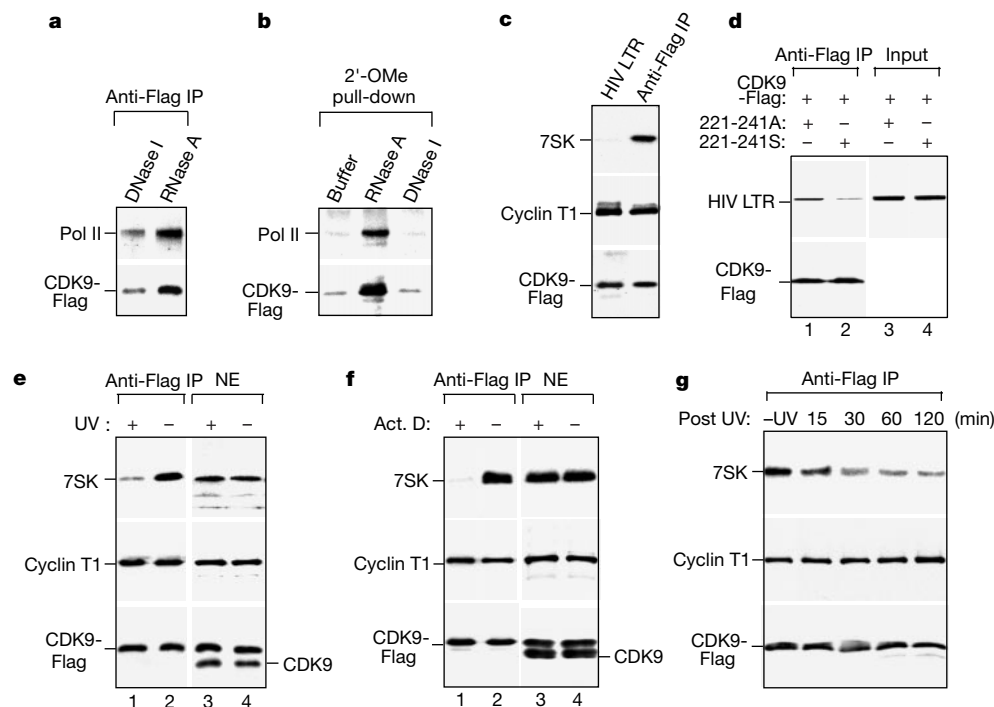


Figure 4 Mechanism of 7SK inhibition of transcription and relief of this inhibition by stress-induced 7SK dissociation from P-TEFb. **a**, CDK9-Flag and its associated factors were affinity purified (anti-Flag IP) from F1C2 nuclear extract, divided into two halves, treated with RNase A or DNase I, and analysed in kinase reactions containing Pol II as a substrate. Phosphoproteins were examined by autoradiography. **b**, The P-TEFb bound to 7SK was isolated from HeLa nuclear extracts (see Methods), and treated and analysed as in **a**. **c**, An immobilized HIV-1 promoter DNA was incubated with F1C2 nuclear extract to assemble PIC¹⁷. The promoter-bound 7SK, cyclin T1 and CDK9-Flag were compared by northern and western analyses with those in total P-TEFb affinity purified from F1C2 nuclear extract. **d**, HeLa cells with an integrated HIV-1 promoter were cotransfected with

CDK9-Flag and the indicated oligonucleotide. Lanes 1 and 2 show the CDK9-Flag-associated HIV-1 promoter DNA as revealed by CHIP analysis (top), and CDK9-Flag in immunoprecipitated chromatin by immunoblotting (bottom). Lanes 3 and 4 contain the HIV-1 DNA amplified from the input chromatin. **e**, CDK9-Flag and its associated factors were affinity purified from nuclear extract (NE) of ultraviolet-irradiated⁸ F1C2 cells, and analysed by northern and immunoblotting. 7SK, cyclin T1, stably transfected CDK9-Flag and the endogenous CDK9 were also detected in F1C2 nuclear extracts. **f**, F1C2 cells were treated with actinomycin D and analysed as in **e**. **g**, F1C2 cells irradiated with ultraviolet were allowed to recover for the indicated time, and were analysed as in **e**. –UV, no UV treatment.

15–30 min after treatment (Fig. 4g), long before any signs of apoptosis appeared. Thus, stress signals such as ultraviolet irradiation and actinomycin D can cause fast and efficient 7SK dissociation from P-TEFb, which may explain their positive effects on HIV-1 transcription and Pol II phosphorylation.

The biological function of 7SK has been elusive since its discovery^{5,22}. We show here that it binds to P-TEFb to inhibit the kinase and transcriptional activity of P-TEFb. Notably, HeLa nuclear extract that was depleted of the P-TEFb bound to 7SK demonstrated no increase in transcription (Fig. 2c), suggesting that the 7SK/P-TEFb complex does not act in a dominant negative fashion over the free P-TEFb in the extract. This may be explained by the inhibitory effect of 7SK on the recruitment of P-TEFb to the PIC. Compared with many other promoters, transcription of the HIV-1 LTR is particularly sensitive to P-TEFb. Consequently, controlling the cellular level of active P-TEFb through the 7SK switch may provide an efficient method to regulate HIV-1 transcription and escape from latency. In support of this proposal, we show here that the 7SK/P-TEFb interaction is disrupted rapidly by the treatment of cells with ultraviolet irradiation and actinomycin D, both of which markedly enhance HIV-1 transcription^{6,8}. Furthermore, ultraviolet irradiation of human T cells before HIV-1 infection significantly shortens the viral growth cycle⁸. Besides stress-induced HIV-1 transcription, dissociation of 7SK from P-TEFb may also allow cells to turn on cellular genes in response to stress. In fact, rapid recruitment of P-TEFb to heat-shock loci has been associated with the activation of heat-shock gene transcription²³. Mammalian cells express several cyclin kinase inhibitors that are critical for cell-cycle

control²⁴. Although CDK9/cyclin T1 has not been implicated in cell-cycle regulation, it nevertheless uses a similar strategy involving an snRNA. Future analyses are necessary to reveal the mechanism of action of 7SK and the stress-induced signalling pathway controlling the binding of 7SK to P-TEFb.

Methods

Constructs and stable CDK9-Flag-expressing cell line

In vitro transcription template pSV40EP-G400 was generated by cloning a 400-nucleotide G-less cassette (G400)¹³ into the *Hind*III and *Eco*RV sites downstream of the SV40 early promoter in pGL-2 (Promega). We performed *in vitro* transcription assay as described¹³. To generate the F1C2 cell line expressing CDK9-Flag, HeLa cells were stably transfected with pBabe-puro-CDK9-Flag, which expresses Flag-tagged CDK9 and confers puromycin resistance. We selected clone F1C2 because CDK9-Flag is expressed at a similar level as the endogenous CDK9. CDK9-Flag and its associated factors were affinity purified from F1C2 nuclear extract using anti-Flag agarose beads (Sigma). After extensive washes with buffer D (20 mM HEPES, pH 7.9, 15% glycerol, 0.2 mM EDTA, 1 mM dithiothreitol, 0.5 mM phenylmethyl sulphonyl fluoride) containing 0.4 M KCl and 0.2% NP40, the materials were eluted by Flag peptide as described⁹.

Transfection of cells with deoxyoligonucleotides

For luciferase assays, HeLa cells were seeded at 2×10^5 cells per well in a 6-well dish one day before transfection. Using the Lipofectamine-Plus Kit (Invitrogen), cells were cotransfected with 100 ng of the indicated luciferase reporter constructs, 2 μ g high-performance liquid chromatography-pure deoxyoligonucleotides, and 20 ng Tat-expressing construct when indicated. Cell lysates were analysed for luciferase activity at 48 h after transfection. For chromatin immunoprecipitation assays (CHIP), a total of 6×10^6 HeLa cells containing an integrated HIV-1 LTR was seeded one day before transfection into five 10-cm dishes. Cells were cotransfected with 12 μ g per dish of the indicated oligonucleotides and 4 μ g per dish of a construct expressing CDK9-Flag. At 36 h after transfection, cells were processed for crosslinking and CHIP with anti-Flag agarose beads as described¹⁹. The HIV-

1 promoter region between -168 and +82 was amplified by polymerase chain reaction from the precipitated chromatin. The sequences of the various deoxyoligonucleotides used in transfection are: 11-31A, 5'-CAGATGTCGACGCCAGATCGC-3'; 11-31S, 5'-CCCCAAAGGGCCCTTAATGGG-3'; 95-114A, 5'-CGCACATGGAGCGGTGAGGGA-3'; 95-114S, 5'-TGCGAGAGTGGACGACGAGC-3'; 221-241A, 5'-CCTTGAGAGCTTGTTTGAGG-3'; 221-241S, 5'-CGTCGATGTGATGCTGTGTGA-3'.

Depletion, purification and cleavage of 7SK

Depletion of the 7SK RNP from HeLa nuclear extract was performed as described⁴ with some modifications. Briefly, 300 µl HeLa extract in buffer D plus 0.1 M KCl, 0.05% NP40 and 0.2 U µl⁻¹ RNasin was incubated at 30 °C for 30 min with 1.8 µM of the biotinylated antisense 2'-O-methyl RNA oligonucleotide (5-biotin-ACCUGAGAGCUUGUUU-GAGG-3') that is complementary to a region in 7SK from nucleotide 221 to 241. The reaction mixture was then incubated for 1 h at 4 °C with streptavidin agarose beads (Sigma). After repeating the procedure twice, the beads were washed with buffer D containing 0.4 M KCl and 0.2% NP40, and the associated 7SK RNP was analysed. To elute 7SK RNP from the beads, we used a 2.5-fold excess displacement deoxyoligonucleotide (5'-CGATATCTCCAAACAAGCTCTCAAGGAGCT-3') in buffer D (0.1 M KCl). We performed oligonucleotide-directed RNase H cleavage of 7SK as described¹⁰.

Received 16 August; accepted 14 September 2001.

- Jones, K. A. Taking a new TAK on Tat transactivation. *Genes Dev.* **11**, 2593-2599 (1997).
- Price, D. H. P-TEFb, a cyclin-dependent kinase controlling elongation by RNA polymerase II. *Mol. Cell Biol.* **20**, 2629-2634 (2000).
- Wei, P., Garber, M. E., Fang, S. M., Fischer, W. H. & Jones, K. A. A novel CDK9-associated C-type cyclin interacts directly with HIV-1 Tat and mediates its high-affinity, loop-specific binding to TAR RNA. *Cell* **92**, 451-462 (1998).
- Wasserman, D. A. & Steitz, J. A. Structural analyses of the 7SK ribonucleoprotein (RNP), the most abundant human small RNP of unknown function. *Mol. Cell Biol.* **11**, 3432-3445 (1991).
- Zieve, G. & Penman, S. Small RNA species of the HeLa cell: metabolism and subcellular localization. *Cell* **8**, 19-31 (1976).
- Cassé, C., Giannoni, F., Nguyen, V. T., Dubois, M. F. & Bensaude, O. The transcriptional inhibitors, actinomycin D and α -amanitin, activate the HIV-1 promoter and favor phosphorylation of the RNA polymerase II C-terminal domain. *J. Biol. Chem.* **274**, 16097-16106 (1999).
- Kumar, S. *et al.* Activation of the HIV-1 long terminal repeat by cytokines and environmental stress requires an active CSBP/p38 MAP kinase. *J. Biol. Chem.* **271**, 30864-30869 (1996).
- Valerie, K. *et al.* Activation of human immunodeficiency virus type 1 by DNA damage in human cells. *Nature* **333**, 78-81 (1988).
- Zhou, Q., Chen, D., Pierstorff, E. & Luo, K. Transcription elongation factor P-TEFb mediates Tat activation of HIV-1 transcription at multiple stages. *EMBO J.* **17**, 3681-3691 (1998).
- Black, D. L., Chabot, B. & Steitz, J. A. U2 as well as U1 small nuclear ribonucleoproteins are involved in premessenger RNA splicing. *Cell* **42**, 737-750 (1985).
- Zieve, G., Benecke, B. J. & Penman, S. Synthesis of two classes of small RNA species *in vivo* and *in vitro*. *Biochemistry* **16**, 4520-4525 (1977).
- Murphy, S. *et al.* DNA sequences complementary to human 7SK RNA show structural similarities to the short mobile elements of the mammalian genome. *J. Mol. Biol.* **177**, 575-590 (1984).
- Zhou, Q. & Sharp, P. A. Novel mechanism and factor for regulation by HIV-1 Tat. *EMBO J.* **14**, 321-328 (1995).
- Luo, Y., Kurz, J., MacAfee, N. & Krause, M. O. C-myc deregulation during transformation induction: involvement of 7SK RNA. *J. Cell. Biochem.* **64**, 313-327 (1997).
- Schnapp, G., Rodi, H. P., Rettig, W. J., Schnapp, A. & Damm, K. One-step affinity purification protocol for human telomerase. *Nucleic Acids Res.* **26**, 3311-3313 (1998).
- Ping, Y. H. & Rana, T. M. Tat-associated kinase (P-TEFb): a component of transcription preinitiation and elongation complexes. *J. Biol. Chem.* **274**, 7399-7404 (1999).
- Zhou, M. *et al.* Tat modifies the activity of CDK9 to phosphorylate serine 5 of the RNA polymerase II carboxyl-terminal domain during human immunodeficiency virus type 1 transcription. *Mol. Cell Biol.* **20**, 5077-5086 (2000).
- Fong, Y. W. & Zhou, Q. Relief of two built-in autoinhibitory mechanisms in P-TEFb is required for assembly of a multicomponent transcription elongation complex at the human immunodeficiency virus type 1 promoter. *Mol. Cell Biol.* **20**, 5897-5907 (2000).
- Boyd, K. E., Wells, J., Gutman, J., Bartley, S. M. & Farnham, P. J. c-Myc target gene specificity is determined by a post-DNA binding mechanism. *Proc. Natl Acad. Sci. USA* **95**, 13887-13892 (1998).
- Kretz-Remy, C. & Arrigo, A. P. The kinetics of HIV-1 long terminal repeat transcriptional activation resemble those of hsp70 promoter in heat-shock treated HeLa cells. *FEBS Lett.* **353**, 339-344 (1994).
- Vlach, J. *et al.* Induction of Sp1 phosphorylation and NF- κ B-independent HIV promoter domain activity in T lymphocytes stimulated by okadaic acid. *Virology* **208**, 753-761 (1995).
- Weinberg, R. A. & Penman, S. Small molecular weight monodisperse nuclear RNA. *J. Mol. Biol.* **38**, 289-304 (1968).
- Lis, J. T., Mason, P., Peng, J., Price, D. H. & Werner, J. P-TEFb kinase recruitment and function at heat shock loci. *Genes Dev.* **14**, 792-803 (2000).
- Peter, M. The regulation of cyclin-dependent kinase inhibitors (CKIs). *Prog. Cell Cycle Res.* **3**, 99-108 (1997).

Acknowledgements

We thank L. Jin for technical assistance, I. von Reis for reagent, O. Bensaude for communicating his work before publication, and Y. Fong and S. Stroschein for critical comments on the manuscript. Supported by grants from the National Institute of Health, the American Cancer Society, the Hellman Faculty Fund and the France-Berkeley Fund to Q.Z.

Correspondence and requests for materials should be addressed to Q.Z. (e-mail: qzhou@uclink4.berkeley.edu).

7SK small nuclear RNA binds to and inhibits the activity of CDK9/cyclin T complexes

Van Trung Nguyen*, Tamás Kiss†, Annemieke A. Michels* & Olivier Bensaude*

* Génomique Moléculaire, UMR 8541 CNRS, Ecole Normale Supérieure, 46 rue d'Ulm, 75230 Paris cedex 05, France

† Laboratoire de Biologie Moléculaire Eucaryote du CNRS, Université Paul Sabatier, 118 route de Narbonne, 31062 Toulouse, France

The transcription of eukaryotic protein-coding genes involves complex regulation of RNA polymerase (Pol) II activity in response to physiological conditions and developmental cues. One element of this regulation involves phosphorylation of the carboxy-terminal domain (CTD) of the largest polymerase subunit by a transcription elongation factor, P-TEFb, which comprises the kinase CDK9 and cyclin T1 or T2 (ref. 1). Here we report that in human HeLa cells more than half of the P-TEFb is sequestered in larger complexes that also contain 7SK RNA, an abundant, small nuclear RNA (snRNA) of hitherto unknown function^{2,3}. P-TEFb and 7SK associate in a specific and reversible manner. In contrast to the smaller P-TEFb complexes, which have a high kinase activity, the larger 7SK/P-TEFb complexes show very weak kinase activity. Inhibition of cellular transcription by chemical agents or ultraviolet irradiation trigger the complete disruption of the P-TEFb/7SK complex, and enhance CDK9 activity. The transcription-dependent interaction of P-TEFb with 7SK may therefore contribute to an important feedback loop modulating the activity of RNA Pol II.

The CTD has a principal role in gene expression and its activity is regulated by multi-site phosphorylation at different steps in the transcription cycle⁴⁻⁷. Investigation of the effect of transcriptional inhibitors has revealed that actinomycin D stimulates the CDK9-dependent phosphorylation of the CTD⁸. We therefore looked for differences in the properties of CDK9 between control and transcriptionally arrested human cells. Complexes of CDK9 and cyclin T with various sizes have been reported in human cell extracts^{9,10}. An extract obtained from untreated HeLa cells was therefore fractionated by ultracentrifugation on a glycerol gradient: CDK9 and cyclin T1 and T2 formed a sediment together in two distinct populations of particles (Fig. 1a). A population of small complexes was found in fractions 3 and 4, as anticipated for the canonical CDK9/cyclin T1 or T2 complex with a predicted relative molecular mass of 130,000 (M_r 130K). The CDK7 and cyclin H subunits of the 120K CDK-activating complex (CAK)¹¹ peaked consistently in the same fractions. However, more than half of the CDK9 and cyclin T1 and T2 molecules were found in fractions 6 and 7, indicating that they migrated in larger complexes. The RNA Pol II core enzyme (550K-600K), as detected by its largest subunit (Rpb1), peaked in fraction 7, and the p62 subunit of the 350K general transcription factor TFIIF¹² predominated in fraction 5. CDK9, cyclin T1 and cyclin T2 therefore are distributed in two distinct families of complexes: small ones (less than 130K) and large ones (400K-500K). The latter contains more than half of the CDK9/cyclin T1 and T2 molecules.

Fractions of the glycerol gradient containing the small (fractions 3 and 4) or large (fractions 6 and 7) complexes of CDK9/cyclin T1 or T2 were pooled. From both of the pooled fractions, a CTD kinase activity was immunoprecipitated with antibodies directed against CDK9 or cyclin T1 (Fig. 1b). This activity was highly sensitive to 5,6-dichloro-1- β -D-ribofuranosyl-benzimidazole (DRB), an inhibitor of CDK9 (ref. 13). Both anti-CDK9 and anti-cyclin T1 antibodies recovered much higher kinase activity from the small

complex fractions than from the large complex ones. For specific activity determinations, an extract from G3H cells expressing a haemagglutinin (HA)-epitope-tagged version of cyclin T1 (ref. 14) was fractionated on a glycerol gradient. After immunoprecipitation with an anti-HA monoclonal antibody, the beads obtained from fractions containing the small complexes showed a 1.5-fold higher CTD kinase activity, although they contained tenfold less CDK9 molecules than the beads obtained from the large complex fractions (Fig. 1c). The kinase activity of CDK9 in large complexes was therefore at least 15-fold weaker than in small ones. As activation of transcription elongation by P-TEFb relies on the CTD kinase activity of CDK9 (refs 1, 15), the large CDK9 complexes represent inactive forms of P-TEFb.

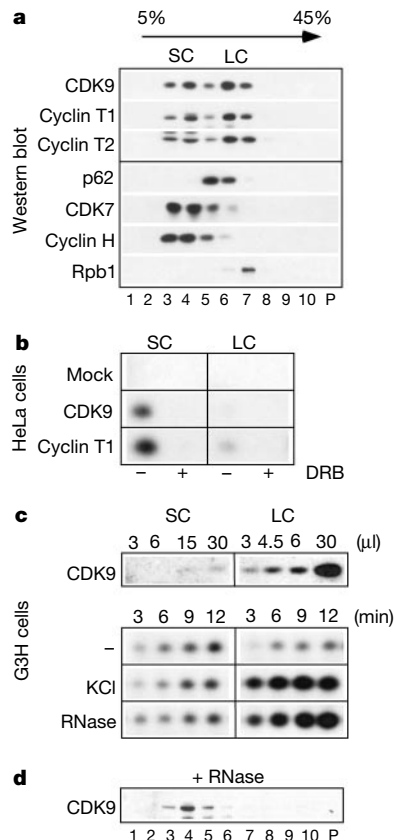


Figure 1 Structural and functional characterization of human CDK9, cyclin T1 and cyclin T2 complexes. **a**, Fractionation of an HeLa cell extract by ultracentrifugation on a glycerol gradient. Distribution of CDK9, cyclin T1 and cyclin T2 in the gradient fractions and pellet (P) was monitored by western blot using specific antibodies. The same blots were probed for Rpb1 (the largest subunit of RNA Pol II), p62 (a TFIIF-specific subunit), CDK7 and cyclin H (subunits of both TFIIF and CAK complexes). SC, small complexes; LC, large complexes. **b**, CTD kinase activities (32 P incorporation into the CTD4 peptide) were immunoprecipitated by anti-CDK9 or anti-cyclin T1 antibodies from pooled HeLa cell gradient fractions 3 and 4 (small complexes, SC) or 6 and 7 (large complexes, LC), and were inhibited by DRB (10 μ M). **c**, Immunoaffinity purification and enzymatic characterization of the small and large CDK9/cyclin T1 complexes. An extract obtained from G3H cells expressing HA-tagged cyclin T1 was fractionated on a glycerol gradient. The CDK9 complexes were immunoprecipitated from pooled fractions 3 and 4 (SC) or fractions 6 and 7 (LC) with the 12CA5 anti-HA monoclonal antibody. The amount of immunoprecipitated CDK9 molecules was determined by western blot analysis and loading various amounts of protein A beads. CTD kinase activities were determined by time-dependent 32 P incorporation into the CTD4 substrate peptide. The immunoprecipitated CDK9/HA-tagged cyclin T1 complexes were either untreated (–) or treated for 5 min at 30 °C with 1 M KCl or with RNase A (10 μ g ml $^{-1}$). **d**, An extract from HeLa cells preincubated with RNase A was fractionated by glycerol gradient centrifugation. CDK9 was monitored by western blot.

On the basis of the results discussed thus far, we proposed that the large form of the human CDK9/cyclin T1 complex might carry an inhibitory factor. To test this assumption, the immunoprecipitated CDK9 complexes were exposed to high concentrations of salt before assaying their CTD kinase activity (Fig. 1c). The salt treatment increased several-fold the CTD kinase activity of the large complex, but did not alter the activity of the small complex. Furthermore, incubation of the immunoprecipitated complexes with ribonuclease A (RNase A) also increased the kinase activity of the large complex. When RNase A or salt was added to the cell extract, only small complexes were recovered in the gradient (Fig. 1d and

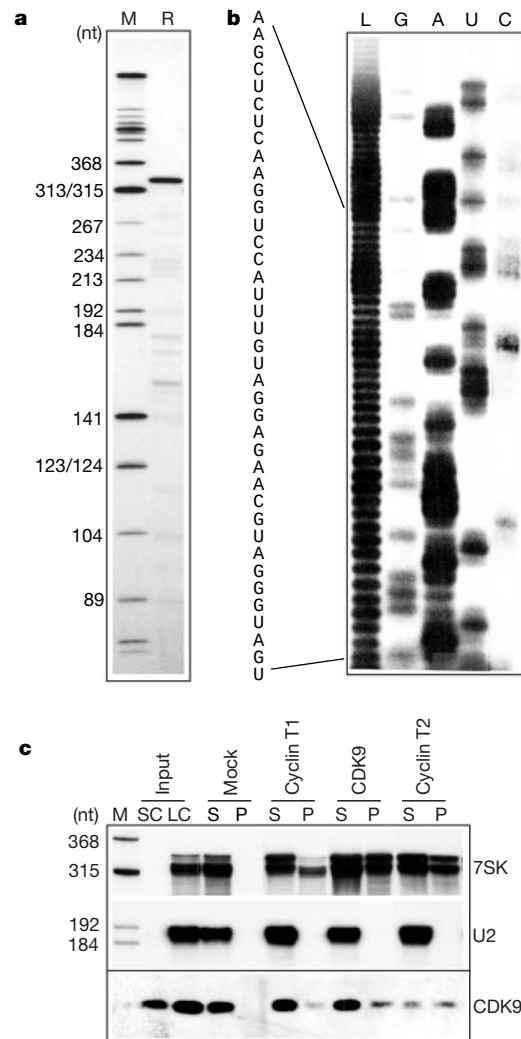


Figure 2 The large complexes of CDK9/cyclin T1 or T2 contain 7SK. **a**, RNAs isolated from large complexes of CDK9/HA-tagged cyclin T1 were labelled at the 3' end by T4 RNA ligase and [32 P]-pCp. The labelled RNAs were size-fractionated on a 6% sequencing gel (R). Size markers (M) were fragments from pBR322 digested by *Hae*III and *Taq*I. **b**, Partial chemical sequencing of the CDK9/HA-cyclin T1 bound RNA species. L, RNA hydrolysed ladder. **c**, The human CDK9 and cyclin T1 and T2 proteins are specifically associated with 7SK. Distribution of 7SK and U2 snRNAs in glycerol gradient fractions of HeLa cell extracts containing the small (SC) and large complexes (LC) of CDK9. The supernatants (S) and bead pellets (P) were obtained by immunoprecipitation from the large gradient fractions with mock, anti-CDK9 and anti-cyclin T1 and T2 antibodies. As HeLa cells contained no HA-tagged proteins, monoclonal antibody 12CA5 was used as a mock antibody. The presence of 7SK and U2 snRNAs was investigated by RNase A/T1 mapping using sequence-specific antisense RNA probes. The appearance of two 7SK RNA-specific protected bands is an artefact, as neither 3'-end labelling (Fig. 2a) nor northern blot analysis (not shown) revealed a size heterogeneity for 7SK. Distribution of CDK9 was monitored by western blot.

data not shown). These results indicate that under *in vitro* conditions the large CDK9 complexes can be converted into small ones that possess a higher CTD kinase activity. The effects of RNase suggested that an RNA-containing inhibitory factor was repressing the CTD kinase activity of the large CDK9/cyclin T complexes.

To identify potential inhibitory RNA(s), the large CDK9/HA-cyclin T1 complex was immunoaffinity purified from G3H cell extracts. RNAs that were recovered from the beads were labelled with cytidine 3',5'-bis(phosphate) (pCp) and analysed on a sequencing gel (Fig. 2a). On exposure, a unique RNA of 330 nucleotides was detected. This RNA was recovered and subjected to chemical sequencing (Fig. 2b). The obtained partial sequence was identical to the previously reported human 7SK sequence from position 150 to 268 (GenBank accession number X05490). At the same time, a full-length complementary DNA of the immunoprecipitated RNA was synthesized, cloned and sequenced. Once again, the obtained 331-nucleotide sequence was identical to that of human 7SK, an abundant snRNA described more than two decades ago². The human 7SK small nuclear ribonucleoprotein particle (RNP) is found in cell extracts as a 12S (400K–500K) RNP complex³. The large CDK9/cyclin T complexes and the previously characterized 7SK snRNP thus show similar sedimentation properties. Indeed, RNase A/T1 mapping experiments demonstrated that 7SK was

present only in those gradient fractions that contained the large CDK9/cyclin T complexes (Fig. 2c). Immunoprecipitation experiments with anti-CDK9, anti-cyclin T1 and anti-cyclin T2 antibodies demonstrated that in contrast to the U2 spliceosomal snRNA, 7SK is specifically associated with all these proteins. These observations establish that CDK9 and cyclin T1 and T2 are subunits of the 7SK snRNP.

We have recently reported that treatment of HeLa cells with actinomycin D markedly stimulates the CDK9-dependent phosphorylation of the CTD⁸. The CTD kinase activity of CDK9, which was immunoprecipitated from an extract of actinomycin-treated cells, showed a 2.5-fold increase compared with untreated cell extracts (Fig. 3a). An increase in CTD kinase activity was also detected after ultraviolet irradiation. Consistently, actinomycin and ultraviolet irradiation promoted the accumulation of small complexes at the expense of the large ones (Fig. 3b). Both treatments favour phosphorylation of RNA Pol II (refs 8, 16) and lead to a general arrest in transcription. DRB, a selective RNA Pol II inhibitor¹⁷, also promoted the disappearance of the CDK9 complexes (Fig. 3b). Removal of DRB from the culture medium allows a prompt recovery of RNA Pol II transcription; the large complexes reappeared when DRB was removed, and the initial distribution of CDK9 was recovered in the gradient. Identical changes in CDK9 distributions were observed when DRB was added and removed in the presence of cycloheximide, which blocks protein synthesis (data not shown). Therefore, neither complex can be considered as an assembly intermediate after synthesis. We instead propose that the two complexes of CDK9 are in dynamic exchange, influenced by the actual activity of RNA Pol II.

The transactivating responsive (TAR)/Tat-dependent recruitment of P-TEFb after initiation of transcription is critical for elongation from the HIV promoter^{1,18}. *In vivo* degradation of 7SK by an antisense oligonucleotide can increase the expression of a cotransfected luciferase reporter gene driven by the HIV promoter (Z. Yang and Q. Zhou, personal communication). Interestingly, stress such as ultraviolet irradiation activates HIV LTR transcription through a TAR/Tat-independent mechanism¹⁹. Furthermore, a moderate inhibition of transcription by actinomycin D or α -amanitin enhances the elongation of transcription initiated on the HIV promoter⁸. Dissociation of 7SK from CDK9 complexes might therefore contribute towards increasing the Tat-independent HIV promoter transcription, a process induced by many stresses that impair overall transcription.

A feedback loop involving protein kinases has been proposed to modulate protein synthesis^{20,21}. Inhibitors of protein synthesis enhance phosphorylation of the S6 ribosomal protein and 4E-BP1, which increases the efficiency of loading messenger RNA on ribosomes. A moderate inhibition favours the translation of a subclass of mRNAs that are restricted at this step^{22,23}. Similarly, the transcription-dependent inhibition of CDK9 activity by 7SK might constitute a feedback loop modulating the efficiency of the elongation step in class II gene transcription. In this context, it has been reported that binding of a small 6S RNA to the *Escherichia coli* RNA polymerase regulates its activity²⁴. Therefore, the RNA-mediated regulation of transcription might be an evolutionarily conserved mechanism. □

Methods

Lysates and gradients

We set up a procedure that ensured complete extraction of P-TEFb subunits. Exponentially growing human HeLa (MRL2 strain) or G3H cells¹⁴ were lysed and vortexed at 4 °C in buffer A (10 mM HEPES, pH 7.9, 150 mM NaCl, 10 mM KCl, 1.5 mM MgCl₂, 1 mM dithiothreitol, 0.5 mM EDTA, 10 U ml⁻¹ RNasin (Amersham), and protease inhibitor cocktail (P8340; Sigma)) supplemented with 0.5 % Nonidet P40. The extracts were clarified by centrifugation for 10 min at 10,000g, and ultracentrifuged for 16 h at 40,000 r.p.m. in a Kontron TST41 rotor at 4 °C on a 5–45% glycerol gradient containing buffer A. The gradients were fractionated into ten samples, and the residual pellets were resuspended in buffer A.

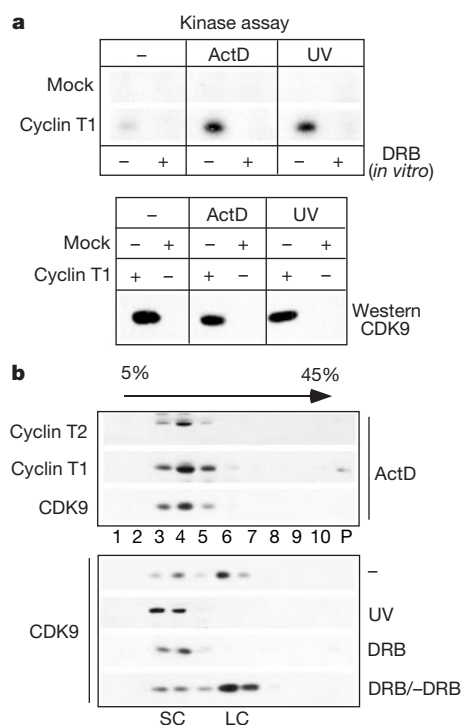


Figure 3 Transcription inhibition promotes disruption of the 7SK/CDK9/cyclin T complexes. **a**, Actinomycin D treatment and ultraviolet irradiation increase the CDK9-dependent *in vitro* phosphorylation of the CTD4 peptide. One hour before lysis, the culture medium was supplemented with 1 μ g ml⁻¹ actinomycin D (ActD) or the cells were irradiated at 254 nm, 40 J m⁻², and allowed to recover for 1 h (UV). Extracts from control or treated HeLa cells were immunoprecipitated by non-immune rabbit serum (mock) or anti-cyclin T1 antibodies, and used for *in vitro* CTD phosphorylation in the presence (+) or absence (-) of DRB (10 μ M). The amount of CDK9 in immunoprecipitates was evaluated by western blot. **b**, Actinomycin D, ultraviolet irradiation and DRB treatment disrupt the large CDK9/cyclin T1 or T2 complexes. HeLa cells were exposed to 100 μ M DRB for 1 h and lysed immediately (DRB). Alternatively, the DRB was removed after 1 h and the cells were allowed to recover for a further 1 h in fresh medium before lysis (DRB/-DRB). The extracts were fractionated on glycerol gradients. Cyclin T1 or T2 and CDK9 were detected in the fractions by western blot.

Antibodies

Rabbit polyclonal anti-Cdk9, anti-cyclin T1 and anti-cyclin T2 were provided by D. Price; mouse monoclonal anti-p62, anti-Cdk7 and anti-cyclin H antibodies were provided by J. M. Egly; monoclonal antibody Pol 3/3 anti-Rpb1 was provided by E. Bautz; and monoclonal antibody 12CA5 anti-HA tag was purchased from BABCo. Antibodies were crosslinked to protein A beads with dimethylpimelidate.

CTD kinase assays

Protein A bead suspensions were incubated at 30 °C with the CTD4 peptide (YSPTSPS)₄ (0.1 µg ml⁻¹) and [γ -³²P]-ATP (100 µM, 0.1 µCi µl⁻¹) with or without DRB (10 µM). The phosphorylated CTD4 peptide was visualized by autoradiography after electrophoresis in a 15% SDS-polyacrylamide gel. The incorporation of ³²P was quantified by phosphorimager.

RNAs

Lysates, gradient fractions or protein A beads were extracted with phenol/guanidinium isothiocyanate and precipitated with isopropanol. The full-length cDNAs of U2 and 7SK snRNAs were cloned into the pBST(KS-) plasmid resulting in pU2 and p7SK recombinant plasmids. After linearization with appropriate restriction enzymes, pU2 and p7SK were used as templates for the synthesis of internally labelled antisense RNA probes by T7 RNA polymerase²⁵. We performed RNase A/T1 mapping as described²⁵.

Received 16 August; accepted 14 September 2001.

- Price, D. H. P-TEFb, a cyclin-dependent kinase controlling elongation by RNA polymerase II. *Mol. Cell. Biol.* **20**, 2629–2634 (2000).
- Zieve, G. & Penman, S. Small RNA species of the HeLa cell: metabolism and subcellular localization. *Cell* **8**, 19–31 (1976).
- Wassarman, D. A. & Steitz, J. A. Structural analyses of the 7SK ribonucleoprotein (RNP), the most abundant human small RNP of unknown function. *Mol. Cell. Biol.* **11**, 3432–3445 (1991).
- Dahmus, M. E. Reversible phosphorylation of the C-terminal domain of RNA polymerase II. *J. Biol. Chem.* **271**, 19009–19012 (1996).
- Corden, J. L. & Patturajan, M. A CTD function linking transcription to splicing. *Trends Biochem. Sci.* **22**, 413–416 (1997).
- Hirose, Y. & Manley, J. L. RNA polymerase II and the integration of nuclear events. *Genes Dev.* **14**, 1415–1429 (2000).
- Bentley, D. Coupling RNA polymerase II transcription with pre-mRNA processing. *Curr. Opin. Cell. Biol.* **11**, 347–351 (1999).
- Cassé, C., Giannoni, F., Nguyen, V. T., Dubois, M.-F. & Bensaude, O. The transcriptional inhibitors, actinomycin D and α -amanitin, activate the HIV-1 promoter and favor phosphorylation of the RNA polymerase II C-terminal domain. *J. Biol. Chem.* **274**, 16097–16106 (1999).
- Garriga, J., Mayol, X. & Grana, X. The CDC2-related kinase PITARE is the catalytic subunit of active multimeric protein complexes. *Biochem. J.* **319**, 293–298 (1996).
- Ramanathan, Y., Reza, S. M., Young, T. M., Mathews, M. B. & Pe'ery, T. Human and rodent transcription elongation factor P-TEFb: interactions with human immunodeficiency virus type 1 Tat and carboxy-terminal domain substrate. *J. Virol.* **73**, 5448–5458 (1999).
- Devault, A. *et al.* MAT1 ('ménagement à trois') a new RING finger protein subunit stabilizing cyclin H-cdk7 complexes in starfish and *Xenopus* CAK. *EMBO J.* **14**, 5027–5036 (1995).
- Svejstrup, J. Q., Vichi, P. & Egly, J.-M. The multiple roles of transcription/repair factor TFIIF. *Trends Biochem. Sci.* **21**, 346–350 (1996).
- Marshall, N. F. & Price, D. H. Purification of P-TEFb, a transcription factor required for the transition into productive elongation. *J. Biol. Chem.* **270**, 12335–12338 (1995).
- O'Keeffe, B., Fong, Y., Chen, D., Zhou, S. & Zhou, Q. Requirement for a kinase-specific chaperone pathway in the production of a Cdk9/cyclin T1 heterodimer responsible for P-TEFb-mediated Tat stimulation of HIV-1 transcription. *J. Biol. Chem.* **275**, 279–287 (2000).
- Yamaguchi, Y. *et al.* NELF, a multisubunit complex containing RD, cooperates with DSIF to repress RNA polymerase II elongation. *Cell* **97**, 41–51 (1999).
- Rockx, D. A. *et al.* UV-induced inhibition of transcription involves repression of transcription initiation and phosphorylation of RNA polymerase II. *Proc. Natl Acad. Sci. USA* **97**, 10503–10508 (2000).
- Egyházi, E. Initiation inhibition and reinitiation of the synthesis of heterogenous nuclear RNA in living cells. *Nature* **262**, 319–321 (1976).
- Garber, M. E. & Jones, K. A. HIV-1 Tat: coping with negative elongation factors. *Curr. Opin. Immunol.* **11**, 460–465 (1999).
- Valerie, K. A. *et al.* Activation of human immunodeficiency virus type 1 by DNA damage in human cells. *Nature* **333**, 78–81 (1988).
- Brown, E. J. & Schreiber, S. L. A signalling pathway to translational control. *Cell* **86**, 517–520 (1996).
- Khaleghpour, K., Pyronnet, S., Gingras, A.-C. & Sonenberg, N. Translational homeostasis: eukaryotic translation initiation factor 4E control of 4E-Binding Protein 1 and p70 S6 kinase activities. *Mol. Cell. Biol.* **19**, 4302–4310 (1999).
- Ignatz, G. G., Hokari, S., DePhilip, R. M., Tsukada, K. & Lieberman, I. Lodish model and regulation of ribosomal protein synthesis by insulin-deficient chick embryo fibroblasts. *Biochemistry* **20**, 2550–2558 (1981).
- Nielsen, F. C., Ostergaard, L., Nielsen, J. & Christiansen, J. Growth-dependent translation of IGF-II mRNA by a rapamycin-sensitive pathway. *Nature* **377**, 358–362 (1995).
- Montzka Wassarman, K. & Storz, G. 6S RNA regulates *E. coli* RNA polymerase activity. *Cell* **101**, 613–623 (2000).
- Goodall, G. J., Wiebauer, K. & Filipowicz, W. Analysis of pre-mRNA processing in transfected plant protoplasts. *Methods Enzymol.* **181**, 148–161 (1990).

Acknowledgements

This research was supported by grants from the Agence Nationale de Recherche sur le Sida,

Association pour la Recherche sur le Cancer, the Ligue Nationale Contre le Cancer (Comité de Paris), the France-Berkeley fund and a Federation of European Biochemical Societies fellowship to A.A.M. We thank P. Ascher, M.-F. Dubois, D. H. Price and all members of the Groupe de Biologie Cellulaire de la Transcription for help and discussions, and Z. Yang and Q. Zhou for communicating their work before publication and providing the G3H cells.

Correspondence and requests for materials should be addressed to O.B. (e-mail: bensaude@wotan.ens.fr).

Mechanism of ubiquitin activation revealed by the structure of a bacterial MoeB–MoaD complex

Michael W. Lake*, Margot M. Wuebbens†, K. V. Rajagopalan† & Hermann Schindelin*

* Department of Biochemistry and Center for Structural Biology, State University of New York at Stony Brook, Stony Brook, New York 11794-5115, USA

† Department of Biochemistry, Duke University Medical Center, Durham, North Carolina 27710, USA

The activation of ubiquitin and related protein modifiers^{1,2} is catalysed by members of the E1 enzyme family that use ATP for the covalent self-attachment of the modifiers to a conserved cysteine. The *Escherichia coli* proteins MoeB and MoaD are involved in molybdenum cofactor (Moco) biosynthesis, an evolutionarily conserved pathway^{3,4}. The MoeB- and E1-catalysed reactions are mechanistically similar, and despite a lack of sequence similarity, MoaD and ubiquitin display the same fold including a conserved carboxy-terminal Gly-Gly motif⁵. Similar to the E1 enzymes, MoeB activates the C terminus of MoaD to form an acyl-adenylate. Subsequently, a sulphurtransferase converts the MoaD acyl-adenylate to a thiocarboxylate that acts as the sulphur donor during Moco biosynthesis^{6,7}. These findings suggest that ubiquitin and E1 are derived from two ancestral genes closely related to *moaD* and *moeB*^{3,5}. Here we present the crystal structures of the MoeB–MoaD complex in its apo, ATP-bound, and MoaD-adenylate forms, and highlight the functional similarities between the MoeB- and E1-substrate complexes. These structures provide a molecular framework for understanding the activation of ubiquitin, Rub, SUMO and the sulphur incorporation step during Moco and thiamine biosynthesis.

The crystal structure (Table 1) of the complex between the *E. coli* MoeB and MoaD proteins reveals a MoeB₂–MoaD₂ heterotetramer (Fig. 1a) in which the MoeB subunits form a dimer. This dimer interface is primarily hydrophobic and buries a surface area of 5,400 Å². To distinguish between the different subunits in the complex, residue numbers are prefixed with either B or D to indicate their location in MoeB or MoaD, respectively. The structure of MoeB consists of eight β -strands, which form a continuous sheet surrounded by eight α -helices. In the amino-terminal half of the sheet, all β -strands are parallel and reveal a variation of the Rossman fold in which the $\beta\alpha\beta\alpha$ -topology near the N terminus is interrupted between the second β -strand and α -helix (β 2 and α 4) by the insertion of two 3_{10} helices. The first 3_{10} helix (3_{10} -A) contains five residues that are strictly conserved in the MoeB/E1 superfamily. A loop between β 1 and α 3 includes a highly conserved glycine-rich motif, Gly-X-Gly-Ala/Gly-Ile/Leu-Gly (where X denotes any residue), reminiscent of the P loop typically found in enzymes that hydrolyse ATP⁸. The C-terminal half of MoeB contains an

Table 1 Crystallographic statistics

Data collection and structure solution					
Parameters	Apo complex	K[Au(CN) ₂]	K ₂ [Pt(CN) ₄]	LuAc ₃	(CH ₃) ₃ PbAc
Resolution (Å)	50–1.7	50–3.0	50–2.4	50–3.0	50–3.5
Wavelength (Å)	1.1	1.1	1.1	1.1	1.1
Number of sites	—	6	4	2	4
<i>R</i> _{sym} (%)	8.7 (40.8)	15.2 (35.4)	9.2 (22.2)	9.0 (21.5)	13.4 (23.0)
Completeness (%)	99.7 (99.3)	99.6 (99.8)	91.9 (67.1)	99.9 (100)	96.9 (99.2)
<i>I</i> / <i>σ</i> (<i>I</i>)	17.5 (1.9)	7.5 (1.9)	10.2 (1.8)	6.6 (1.9)	5.8 (2.2)
Phasing power	—	1.17/1.33	1.01/0.96	0.68/0.65	0.92/1.04
<i>R</i> _{Cullis}	—	0.80/0.83	0.84/0.85	0.89/0.88	0.89/0.90
FOM	—	0.414/0.343			

Refinement statistics

Parameters	Apo complex	ATP	Acyl-adenylate
Resolution (Å)	20–1.7	20–2.9	20–2.1
Number of reflections	36,267	6,696	17,376
Protein/cofactor/solvent atoms	2,408/0/231	2,238/31/0	2,404/23/140
Data to parameter ratio	3.4	0.7	1.7
<i>R</i> (<i>R</i> _{free})	0.175 (0.208)	0.203 (0.267)	0.187 (0.224)
Ramachandran statistics (%)	90.7/8.2/0.7/0.4	85.4/13.5/0.8/0.4	88.8/10.1/0.7/0.4
Overall average B-factor (Å ²)	24.9	49.7	32.3

$R_{\text{sym}} = \sum_i \sum_j |I_i - \langle I \rangle| / \sum_i \sum_j I_i$, where I_i is the i th measurement and $\langle I \rangle$ is the weighted mean. $\langle I \rangle / \sigma(I)$ indicates the average of the intensity divided by its standard deviation. Numbers in parenthesis refer to the respective highest resolution shells. Phasing power is the mean value of the heavy atom structure factor amplitude divided by the lack of closure for isomorphous/anomalous differences. R_{Cullis} is the lack of closure divided by the absolute of the difference between F_{PH} and F_P for isomorphous differences of acentric/centric data. FOM is the figure of merit for acentric/centric data. $R_{\text{cryst}} = \sum_i |F_o| - |F_c| / \sum_i |F_o|$ where F_o and F_c are the observed and calculated structure factor amplitudes. R_{free} is the same as R_{cryst} for 5% of the data randomly omitted from refinement. Ramachandran statistics indicate the fraction of residues in the most favoured, additionally allowed, generously allowed and disallowed regions of the Ramachandran diagram²⁶.

antiparallel β -sheet ($\beta 5$ – $\beta 8$) in a fold distantly related to a family of sugar-binding proteins.

The MoeB–MoaD interface is 65% hydrophobic in character, a situation similar to that observed in molybdopterin synthase⁵. This protein, a complex of MoaD and another essential subunit, MoaE, is responsible for incorporating the sulphur of the MoaD thiocarboxylate into the dithiolene group of Moco⁷. In both complexes, the hydrophobic core involves the region of MoaD containing Phe D7, Leu D59 and Phe D75 (Fig. 1b). Most of the MoaD-binding interactions are localized to $\alpha 7$, $\beta 7$ and $\beta 8$ of MoeB. In addition to the hydrophobic core, hydrogen bonds and ionic interactions are present between the side chains of Asp B227 and Arg D11. Interestingly, Asp B227 is replaced by glutamate, glutamine and asparagine in members of the E1 family, and Arg D11 is replaced by lysine in ubiquitin, possibly representing the conservation of a critical interaction in these complexes. The hydrophobic surface of MoaD

involved in MoeB binding is partially conserved in ubiquitin; however, two ubiquitin surface arginines (Arg42 and Arg72) involved in E1 binding⁹ are absent from MoaD, indicating that the interactions between ubiquitin and E1 differ to some extent. The most notable feature of the MoeB–MoaD interface is the C-terminal extension of the MoaD C terminus (residues 76–81) into a pocket on the MoeB surface (Fig. 1b). The C terminus of MoaD extends over $\beta 5$ of MoeB, which acts as a structural scaffold. Sequence alignments of MoeB and E1 show a preference for small amino acids (Gly, Ala, Ser) at the centre of $\beta 5$, facilitating the insertion of the Gly–Gly motifs of MoaD and ubiquitin into the active sites of MoeB and E1.

In the MoeB–MoaD–ATP ternary complex (Table 1), ATP is bound in close proximity to the C terminus of MoaD (Fig. 2a), with residues in the P loop forming the floor of the nucleotide-binding pocket, and the adenine ring nonspecifically bound in a hydrophobic patch. ATP is anchored at the active site by its triphosphate moiety and ribose hydroxyls. The α -phosphate is buried deeply in the pocket and forms main-chain contacts with Gly B41 of the P loop. The strictly conserved Arg B73 contacts one oxygen each of the α - and β -phosphates. Lys B86 also interacts with the β -phosphate while Ser B69 and Asn B70 anchor the γ -phosphate. The overall shape of the binding pocket distorts the ATP molecule and induces a kink at the α -phosphate. The side chain of Arg B14' from the second MoeB monomer undergoes a significant conformational change compared with the nucleotide-free structure and is within hydrogen-bonding distance of the γ -phosphate. Although the MoaD carboxylate and the ATP α -phosphate are in close spatial proximity, nucleophilic attack of the carboxylate seems to be precluded by electrostatic repulsion between the two negatively charged groups.

Soaking experiments with ATP and Mg^{2+} revealed a MoaD acyl-adenylate intermediate (Table 1) consisting of Gly D81 covalently linked to the α -phosphate through a mixed anhydride (Fig. 2b). Although the pyrophosphate-leaving group is not visible, a bound sulphate molecule from the mother liquor mimics one of the pyrophosphate phosphates. In contrast to glycyl-transfer RNA synthetase, where the metal remains bound to the α -phosphate after formation of the glycyl-adenylate¹⁰, this structure does not contain a bound Mg^{2+} . Comparison of the apo complex with its ATP-bound and acyl-adenylate forms reveals only subtle conformational changes that are localized to the immediate vicinity of the active site. Other than the conformation of the MoaD C terminus where Gly D80 and Gly D81 adopt clearly different conformations on acyl-adenylate formation, the active sites of the apo and

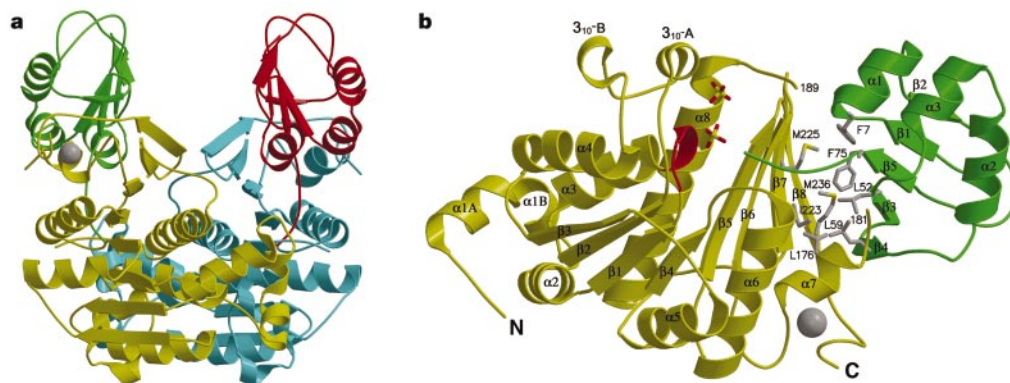


Figure 1 Structure of the MoeB–MoaD complex. **a**, Ribbon diagram of the heterotetramer with MoaD in green and red and MoeB in yellow and cyan. The Zn^{2+} ions (grey spheres, but one is completely hidden) are coordinated with tetrahedral geometry by four cysteines originating from two Cys–X–Cys motifs. All residues (1–81) of MoaD and residues 2–181 and 189–248 of MoeB are shown. **b**, Hydrophobic interactions between MoeB (yellow) and MoaD (green). The glycine-rich P loop motif of MoeB is highlighted in red; two

sulphate molecules bound in the apo complex are shown in all-bonds representation. One sulphate is ligated by strictly conserved residues in helix 3₁₀–A and the other is in close proximity to the MoaD Gly–Gly motif. Secondary structure elements, terminal residues and those adjacent to the disordered loop are labelled. Figs 1 and 2a, b were generated with Molscript²⁶.

acyl-adenylate models are markedly similar. In contrast, the ATP-bound model shows the most pronounced structural changes, particularly in the side chain of Arg B14' and the region around Asp 130 (Fig. 2a, b).

The three MoeB–MoaD structures provide a dynamic view of the MoeB active site and have allowed us to propose a reaction mechanism for MoaD activation (Fig. 2c). Specifically, while the conserved Asp 130 is predicted to ligate the Mg^{2+} ion, thereby alleviating the electrostatic repulsion between the C terminus of MoaD and the α -phosphate of ATP, a carboxylate oxygen of Gly D81 attacks the α -phosphate to create a transient, penta-coordinated intermediate. The structural changes around Asp 130 in the ATP complex could thus be a consequence of the absence of the divalent metal. The strained conformation of the triphosphate observed in the ATP-bound structure would facilitate subsequent cleavage of the bond between the α - and β -phosphates. The conserved residues in helix 3₁₀-A (particularly Arg B73) and Arg B14' of the second MoeB monomer would stabilize the developing negative charge on the β -phosphate. At this point, the pathways for Moco biosynthesis and ubiquitylation diverge to produce either thiocarboxylated MoaD or the ubiquitin–E1 thioester complex.

To verify the importance of the principal catalytic residues in this proposed reaction mechanism, MoeB variants were generated with substitutions at residues Arg 14, Arg 73 and Asp 130. The ability of these variants to support Moco biosynthesis in a *moeB*[−] *E. coli* strain was assessed with an overlay assay specific for the activity of the Moco-containing enzyme, nitrate reductase (Fig. 2d). Compared with wild-type MoeB, complementation with the Asp130Ala variant failed to produce active nitrate reductase and the Asp130Glu variant

had substantially reduced activity. The Arg14Lys, Arg14Ala and Arg73Ala variants were only slightly impaired, indicating that these arginines can at least partially compensate for each other. Consistent with this assumption, the Arg14Ala and Arg73Ala double mutant was completely inactive, and the Arg14Lys and Arg73Ala variant had significantly reduced activity. These results demonstrate the requirement for at least one positively charged residue in the vicinity of the β / γ -phosphate moiety of ATP. However, the presence of two positive charges does not necessarily result in full catalytic activity, as indicated by substantially reduced activity for the Arg73Lys variant. These studies verify that both Arg 14 and Arg 73 are crucial for MoeB activity and suggest an intricate interplay between the side chains of these residues during the catalytic cycle.

Sequence alignments show that Cys B187, a residue located in the disordered loop region in close proximity to the active site, corresponds to the conserved, active site cysteine in the E1 family¹¹. It has been proposed that Cys 187 forms a thioester with the C terminus of MoaD that is then attacked by a sulphur atom to form the thiocarboxylate². However, a recent study of the interaction between MoaD and MoeB found no evidence for formation of such a thioester and no loss of activity for the Cys187Ala variant⁶. Activation of a small ubiquitin-like protein is also seen during thiamine biosynthesis, where ThiF and ThiS display sequence homology with MoeB and MoaD, respectively. For thiamine biosynthesis, recent findings have implicated the NifS-like protein, IscS, in mobilization of sulphur from cysteine to form an IscS-persulphide that is subsequently shuttled to ThiI (a rhodanese-like enzyme) to form a putative ThiI-persulphide^{12,13}. This moiety is responsible for attacking the adenylated ThiS bound at the ThiF

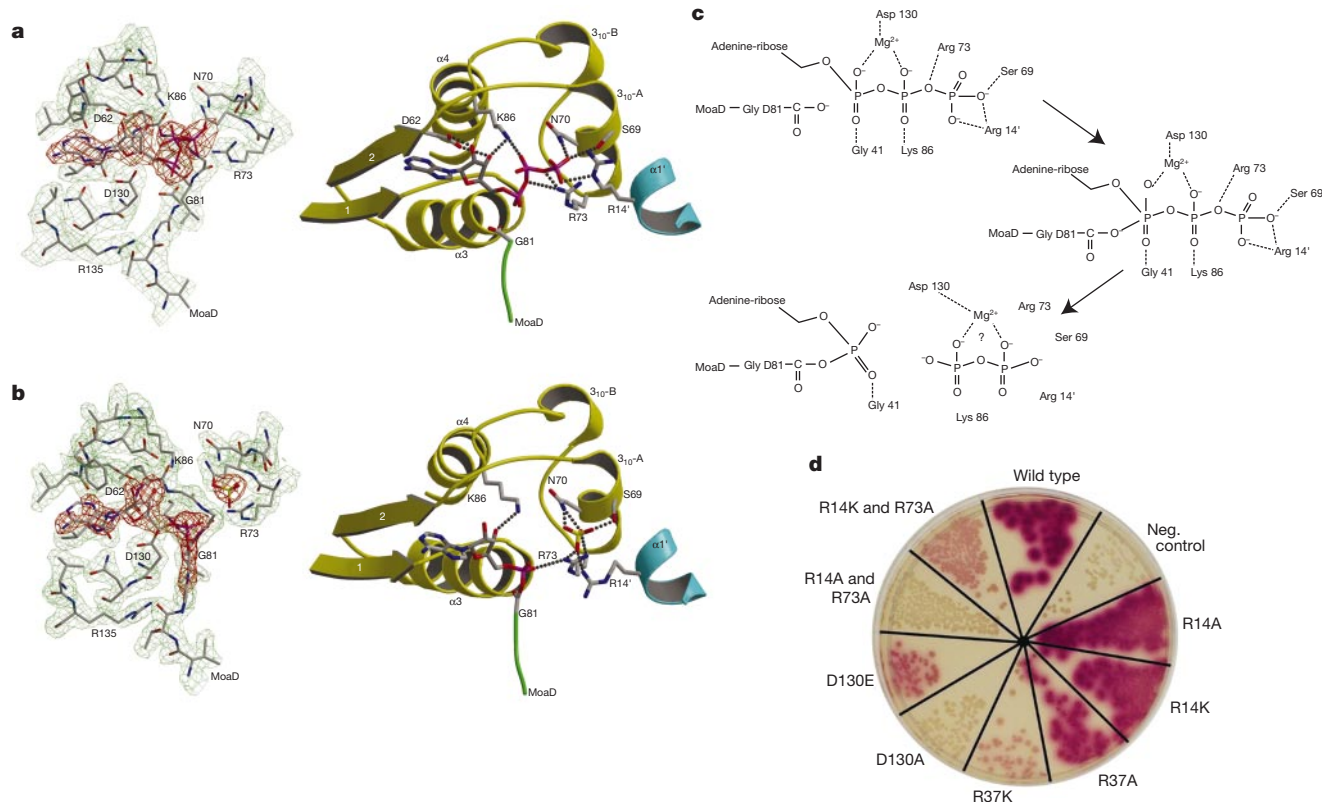


Figure 2 MoeB-catalysed activation of MoaD. **a**, Electron density maps (left) and ribbon diagram (right) of the ATP complex. A $2F_o - F_c$ electron density map (green, 1σ) encompasses residues from the MoeB active site and MoaD C terminus. A $F_o - F_c$ electron density map (red, 3σ) shows the ATP. Arg B14' is shown in cyan in the ribbon diagram. **b**, Electron density maps (left), as described in **a**, and ribbon diagram (right) of the MoeB–MoaD acyl-adenylate. The difference electron density map covers the covalently

bound acyl-adenylate, a sulphate molecule and Gly D81. **c**, Postulated reaction mechanism of MoaD-adenylate formation. The Mg^{2+} may remain associated with the pyrophosphate-leaving group. **d**, Nitrate reductase overlay assay of *E. coli* MoaB- cells expressing either wild-type or MoeB variants. For the negative control, cells were transformed with the pET15b vector alone.

active site, leading to the formation of thiocarboxylated ThiS through a ThiF/ThiS acyl-disulphide intermediate¹⁴. Several NifS-like proteins have also been implicated in cysteine sulphur mobilization and transfer during Moco biosynthesis¹⁵. Notably, some MoeB orthologues, including human MoeB, have a C-terminal extension that contains distant sequence relationships with rhodanases. Together, these observations suggest that sulphur incorporation during Moco and thiamine biosynthesis is mechanistically similar, but that the importance of the residue corresponding to the active site cysteine of E1 is controversial^{6,14}. The disorder of the surface loop that includes Cys B187 in all three MoeB–MoaD structures could indicate that these residues are involved in protein–protein interactions with either a rhodanase-like protein or a cysteine desulphurase.

Multiple sequence alignments of MoeB and different members of the E1 family including the ubiquitin-, Rub- and SUMO-activating enzymes (Fig. 3a, b) reveal a remarkable degree of conservation for the residues surrounding the active site (Fig. 3c). On the basis of the structural evidence presented here, it is possible to assign functional roles to most of these residues. The loop region between β_1 and α_3 (yellow) consists of a glycine-rich, nucleotide-binding motif that facilitates ATP entry into the active site. The loop region between β_2 and helix 3_{10} -A (red) is critical for binding the ribose. The highly conserved residues forming helix 3_{10} -A (cyan) are essential for binding the β - and γ -phosphates of ATP and, more importantly, stabilizing the pyrophosphate-leaving group on attack by the MoaD- or ubiquitin-carboxylates. Residues in the loop between β_4 and α_6 (green) are responsible for proper positioning of Asp B130 (predicted to be involved in Mg^{2+} ligation) adjacent to the α -phosphate. In this same region, Arg B135 inside helix α_6 properly orients both the incoming C-terminal extension of MoaD and strand β_5 of MoeB, which serves to support the C-terminal MoaD Gly-Gly dipeptide.

In light of the sequence homologies, our studies suggest that enzymes involved in the activation of ubiquitin, Rub, SUMO and ThiS all contain a structurally similar, MoeB-like domain. Careful sequence analysis reveals that the ubiquitin-activating enzymes contain an additional MoeB-like domain near their N terminus

(Fig. 3a). Although most of the residues in the signature sequence motifs required for ATP-binding and hydrolysis (Asn 70, Arg 73, Lys 86, Asp 130 and Arg 135) are missing from the first MoeB-like domain, the residue corresponding to Arg 14 is strictly conserved. As described above, Arg 14 is inserted into the active site across the dimer interface and has a critical role during ATP hydrolysis. Although the enzymes involved in the activation of SUMO and Rub contain only a single MoeB-like domain, there are additional enzymes in each of these pathways (AOS1 and AXR1) that are essential for catalysis and also contain a MoeB-like domain with an Arg residue corresponding to Arg 14 (Fig. 3b). These results strongly suggest that this group of proteins mimics the dimeric structure of MoeB by arranging two MoeB-like domains, present on the same or two different polypeptide chains, in a manner similar to that observed here in the MoeB dimer. In contrast to the two active sites of the MoeB dimer, the monomeric forms of these enzymes are predicted to contain one active site created by residues originating from both MoeB-like domains. The results presented here reveal that although members of the MoeB/E1 enzyme superfamily have diversified to be used in seemingly unrelated pathways, their mechanism for acyl-adenylate formation has been evolutionarily conserved. □

Methods

Crystallization and structure determination

MoeB was purified as described⁶, and MoaD purification will be described elsewhere. Crystals of the MoeB–MoaD complex were grown by incubating equal volumes of 23 mg ml⁻¹ MoeB and 10 mg ml⁻¹ MoaD at 4 °C for 1 h, followed by hanging-drop vapour diffusion against a reservoir containing 1.7 M Li₂SO₄ and 100 mM HEPES, pH 7.5. Crystals belonged to the tetragonal space group *P*4₁2₁2 with unit cell dimensions of *a* = 77.8 Å and *c* = 102.1 Å. Before cryo-cooling, crystals were transferred into a solution consisting of 1.5 M Li₂SO₄ and 25% glycerol. We collected all diffraction data on beamline X26C of the National Synchrotron Light Source at Brookhaven National Laboratory. Heavy atom derivatives were prepared by soaking crystals in 2 mM K₂[Pt(CN)₄] for 24 h, 2 mM K[Au(CN)₂] for 24 h, 10 mM Lu-acetate for 72 h, or 10 mM (CH₃)₃Pb-acetate for 72 h. All diffraction data were indexed, integrated and scaled with the HKL software¹⁶. Heavy atom sites were detected by direct methods using SHELX¹⁷ and difference Fourier techniques. Phase refinement was performed with SHARP¹⁸ at 3 Å resolution followed by solvent flattening and gradual phase extension to 1.7 Å with SOLOMON¹⁹. The resulting electron density map was used to autobuild the polypeptide chain using amplitudes to

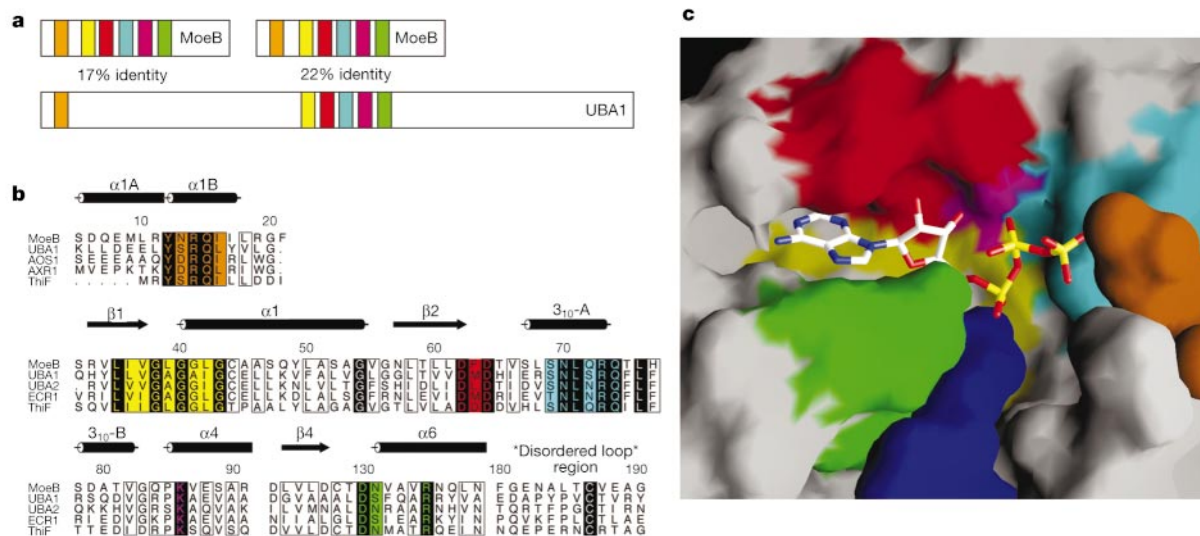


Figure 3 Active site conservation in the MoeB/E1 enzyme superfamily. **a**, Schematic diagram of sequence relationships between MoeB and UBA1, with conserved sequence motifs indicated by coloured blocks. *E. coli* MoeB and residues 1–250 and 400–660 of human UBA1 are 17% and 22% identical, respectively. **b**, Excerpts from a multiple sequence alignment of MoeB/E1 superfamily enzymes. The alignment²⁷ is based on several representatives of each group, but only one family member is displayed: MoeB (*E. coli*), UBA1 (human ubiquitin-activating enzyme), AOS1/UBA2 (heterodimeric human

SUMO-activating enzyme), AXR1/ECR1 (heterodimeric *Arabidopsis thaliana* Rub-activating enzyme) and ThiF (*E. coli*). The strictly conserved cysteine corresponding to MoeB Cys 187 is also included. **c**, Surface representation²⁸ of the MoeB–MoaD complex around the active site. The surface is colour-coded to correspond to the conserved regions shown in **a** and **b**, with the MoaD C terminus in blue. The bound ATP molecule is shown in all-bonds representation.

1.7 Å with ARP²⁰ after 5% of the data had been set aside to calculate the free R-factor. Additional calculations were performed with the CCP4 suite of programs²¹. The model was refined with REFMAC²² and water molecules were added with ARP. Model building was performed using the program O²³. The final model has been refined at 1.7 Å to an R-factor of 0.175 with an R_{free} of 0.208 (Table 1).

ATP and acyl-adenylate complexes

Crystals of the apo complex were soaked for 24 h in a solution consisting of 1.7 M Li₂SO₄, 100 mM HEPES, pH 7.5, and 20 mM ATP. Diffraction data were collected at beamline X26C (99.7% complete, 6.3-fold redundancy, $R_{\text{sym}} = 12.9\%$), and the structure was solved using difference Fourier methods. The ATP-bound model has been refined at 2.9 Å to an R-factor of 0.203 ($R_{\text{free}} = 0.267$). No water molecules were added owing to the limited resolution. Residues 167–188 and 241–248 appear to be disordered, including the zinc-binding motifs and the previously missing surface loop of MoeB. In the apo complex, the polypeptide segments containing the zinc-binding motifs have higher average B-factors compared with the remainder of the molecule. At 2.9 Å the quality of the electron density maps for the ATP complex is reduced making it impossible to observe these more mobile regions. To obtain the acyl-adenylate complex, apo crystals were soaked for 24 h in a solution consisting of 1.7 M Li₂SO₄, 100 mM HEPES, pH 7.5, 20 mM ATP, and 20 mM MgSO₄. Diffraction data were collected at beamline X26C (98.7% complete, 3.5-fold redundancy, $R_{\text{sym}} = 8.6\%$), and the structure was solved using difference Fourier methods. The acyl-adenylate model has been refined at 2.1 Å resolution to an R-factor of 0.188 ($R_{\text{free}} = 0.225$) following the same protocol described for the apo complex. The residues in the zinc-binding motif of MoeB are present in the electron density maps in a conformation identical to that found in the apo complex, but residues 182–188 are again disordered.

MoeB variants and activity measurements

The QuickChange kit from STRATAGENE was used to generate the Arg14Ala, Arg14Lys, Arg73Ala, Arg73Lys, Asp130Ala and Asp130Glu variants of MoeB as well as the double mutants Arg14Ala and Arg73Ala, and Arg14Lys and Arg73Ala. Nucleic acid sequences were verified by automated sequencing of both strands. *MoeB*[−] cells⁵ transformed with plasmids¹⁵ expressing either the wild-type or a mutated version of MoeB were grown aerobically for 16 h at 37 °C on Luria Broth plates. After anaerobic growth for several hours at 22 °C, an overlay assay for nitrate reductase activity was performed as described²⁴.

Received 19 July; accepted 11 September 2001.

- Hochstrasser, M. All in the ubiquitin family. *Science* **289**, 563–564 (2000).
- Hochstrasser, M. Evolution and function of ubiquitin-like protein-conjugation systems. *Nature Cell Biol.* **2**, E153–E157 (2000).
- Rajagopalan, K. V. in *Escherichia coli and Salmonella typhimurium: Cellular and Molecular Biology* (ed. Neidhardt, F. C.) 674–679 (ASM Press, Washington DC, 1996).
- Rajagopalan, K. V. Biosynthesis and processing of the molybdenum cofactors. *Biochem. Soc. Trans.* **25**, 757–761 (1997).
- Rudolph, M. J., Wuebbens, M. M., Rajagopalan, K. V. & Schindelin, H. Crystal structure of molybdopterin synthase and its evolutionary relationship to ubiquitin activation. *Nature Struct. Biol.* **8**, 42–46 (2001).
- Leimkühler, S., Wuebbens, M. M. & Rajagopalan, K. V. Characterization of *Escherichia coli* MoeB and its involvement in the activation of MPT synthase for the biosynthesis of the molybdenum cofactor. *J. Biol. Chem.* **276**, 34695–34701 (2001).
- Pitterle, D. M., Johnson, J. L. & Rajagopalan, K. V. *In vitro* synthesis of molybdopterin from precursor Z using purified converting factor. *J. Biol. Chem.* **268**, 13506–13509 (1993).
- Walker, J. E., Saraste, M., Runswick, M. J. & Gay, N. J. Distantly related sequences in the alpha- and beta-subunits of ATP synthase, myosin, kinases and other ATP-requiring enzymes and a common nucleotide binding fold. *EMBO J.* **1**, 945–951 (1982).
- Burch, T. J. & Haas, A. L. Site-directed mutagenesis of ubiquitin. Differential roles for arginine in the interaction with ubiquitin-activating enzyme. *Biochemistry* **33**, 7300–7308 (1994).
- Arnez, J. G., Dock-Bregeon, A. C. & Moras, D. Glycyl-tRNA synthetase uses a negatively charged pit for specific recognition and activation of glycine. *J. Mol. Biol.* **286**, 1449–1459 (1999).
- Hatfield, P. M. & Vierstra, R. D. Multiple forms of ubiquitin-activating enzyme E1 from wheat. *J. Biol. Chem.* **267**, 14799–14803 (1992).
- Lauhon, C. T. & Kambampati, R. The *iscS* gene in *Escherichia coli* is required for the biosynthesis of 4-thiouridine, thiamin, and NAD. *J. Biol. Chem.* **275**, 20096–20103 (2000).
- Palenchar, P. M., Buck, C. J., Cheng, H., Larson, T. J. & Mueller, E. G. Evidence that ThiI, an enzyme shared between thiamin and 4-thiouridine biosynthesis, may be a sulfurtransferase that proceeds through a persulfide intermediate. *J. Biol. Chem.* **275**, 8283–8286 (2000).
- Xi, J., Ge, Y., Kinsland, C., McLafferty, F. W. & Begley, T. P. Biosynthesis of the thiazole moiety of thiamin in *Escherichia coli*: identification of an acylsulfide-linked protein-protein conjugate that is functionally analogous to the ubiquitin/E1 complex. *Proc. Natl Acad. Sci. USA* **98**, 8513–8518 (2001).
- Leimkühler, S. & Rajagopalan, K. V. An *Escherichia coli* NifS-like sulfurtransferase is required for the transfer of cysteine sulfur in the *in vitro* synthesis of molybdopterin from precursor Z. *J. Biol. Chem.* **276**, 22024–22031 (2001).
- Otwinowski, Z. & Minor, W. in *Methods in Enzymology: Macromolecular Crystallography* (eds Carter, C. W. & Sweet, R. M.) 307–326 (Academic, San Diego, 1997).
- Sheldrick, G. M. & Schneider, T. R. in *Methods in Enzymology: Macromolecular Crystallography* (eds Carter, C. W. & Sweet, R. M.) 319–343 (Academic, San Diego, 1997).
- DeLaFortelle, E. & Bricogne, G. in *Methods in Enzymology: Macromolecular Crystallography* (eds Carter, C. W. & Sweet, R. M.) 472–494 (Academic, San Diego, 1997).
- Abrahams, J. P. & Leslie, A. G. W. Methods used in the structure determination of bovine mitochondrial F₁ ATPase. *Acta Crystallogr. D* **52**, 30–42 (1996).
- Perrakis, A., Morris, R. & Lamzin, V. S. Automated protein model building combined with iterative structure refinement. *Nature Struct. Biol.* **6**, 458–463 (1999).
- Bailey, S. The CCP4 suite: programs for protein crystallography. *Acta Crystallogr. D* **50**, 760–763 (1994).

- Murshudov, G., Vagin, A. & Dodson, E. Refinement of macromolecular structures by the maximum likelihood method. *Acta Crystallogr. D* **53**, 240–255 (1997).
- Jones, T. A., Zou, J. Y., Cowan, S. W. & Kjeldgaard, M. Improved methods for building protein models in electron density maps and the location of errors in these models. *Acta Crystallogr. A* **47**, 110–119 (1991).
- Johnson, M. E. & Rajagopalan, K. V. *In vitro* system for molybdopterin biosynthesis. *J. Bacteriol.* **169**, 110–116 (1987).
- Laskowski, R. A., Moss, D. S. & Thornton, J. M. Main-chain bond lengths and bond angles in protein structures. *J. Mol. Biol.* **231**, 1049–1067 (1993).
- Kraulis, P. J. MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. *J. Appl. Crystallogr.* **24**, 946–950 (1991).
- Barton, G. J. ALSCRIPT: a tool to format multiple sequence alignments. *Protein Eng.* **6**, 37–40 (1993).
- Nicholls, A., Sharp, K. A. & Honig, B. Protein folding and association: insights from the interfacial and thermodynamic properties of hydrocarbons. *Proteins* **11**, 281–296 (1991).

Acknowledgements

We thank M. J. Rudolph for initial help with crystallization and data collection, J. Daniels for technical assistance, and D. Schneider for support at beamline X26C. This work was supported by National Institutes of Health (NIH) grants to H.S. and K.V.R.. The National Synchrotron Light Source in Brookhaven is supported by DOE and NIH, and beamline X26C is supported in part by the State University of New York at Stony Brook and its Research Foundation.

Correspondence and requests for materials should be addressed to H.S. (e-mail: hermann.schindelin@sunysb.edu). The atomic coordinates have been deposited in the Protein Data Bank under accessions numbers 1JW9, 1JWA and 1JWB.

correction

Antibacterial agents based on the cyclic D,L-α-peptide architecture

Sara Fernandez-Lopez, Hui-Sun Kim, Ellen C. Choi, Mercedes Delgado, Juan R. Granja, Alisher Khasanov, Karin Kraehenbuehl, Georgina Long, Dana A. Weinberger, Keith M. Wilcoxon & M. Reza Ghadiri

Nature **412**, 452–455 (2001).

The two organisms *Streptococcus pneumoniae* and *Enterococcus faecalis* are mislabelled as Gram-negative strains (paragraph 3 and Table 1), whereas they are Gram-positive. This error does not alter our conclusions. □

erratum

Transmission intensity and impact of control policies on the foot and mouth epidemic in Great Britain

Neil M. Ferguson, Christl A. Donnelly & Roy M. Anderson

Nature **413**, 542–548 (2001).

In this Letter, the key to Figure 4b contained two errors. The description of the green curve should be “no non-IP culling” (not “non-IP culling”) and the description of the orange curve should be “69% case increase” (not “69% case decrease”). □

Cultural diversity

Ready-primed keratinocytes, new media and new vessels.

OptiCell

BioCrystall www.biocrystal.com

A new approach to cell culture

OptiCell is a sterile, sealed-cell environment between two optically clear gas-permeable growth surfaces in a standard microtitre plate-sized frame. It is fitted with self-sealing ports to provide aseptic access to the interior and contents. OptiCell is suitable for most cell culture activities, including magnetic cell separation, staining, transfection and cloning. Media volume is 10 ml, and each OptiCell provides 100 cm² of growth surface area. Optically clear growth surfaces allow unimpeded visual access to the growing cells.

Keratinocytes

GIBCO www.invitrogen.com

Keratinocytes, ready to roll

GIBCO Primary Human Keratinocytes are processed and frozen in GIBCO Defined Keratinocyte-SFM, so are effectively adapted to the medium. When recovered from cryopreservation, the cells establish fully confluent cultures within 5 days when stored, thawed and cultured. The lifespan of these keratinocytes is performance-tested for 16 population doublings, with most lots demonstrating more than 20 doublings.

EasY Flask

Nalge Nunc www.nalgenunc.com

Extending the range

EasY Flask cell culture vessels, from Nalge Nunc International, are available with surface areas of 25, 75 and now 175 cm². The flasks are in-line tested and certified non-pyrogenic. The 'Easy' closure fitting is available either as a filter cap or vent/close cap, with a 'Y'-shaped mark on the vent cap for visual verification of the closure position. The Easy Flask design allows air flow under the flask to optimize temperature equilibration during incubation even when stacked, and provides both printed and moulded volume graduations on the flask sides for easy measurement of contents.

Stericup-VP

Millipore www.millipore.com

Removing mycoplasma from culture media

The Stericup-VP vacuum filter cup provides high-speed filtration of up to 1 litre with low protein binding, and also features the 0.1 µm Millipore Express membrane to remove mycoplasma. The membrane's asymmetric structure is designed to offer high filtration capacity and speed. The unit is recommend-



Manifold delight: NaviCyte in close-up.

ed for use in laminar flow hoods. Its new no tip/easy grip flask design and compact profile are claimed to improve stability during filtration and make gripping the receiver easier. The bottom of the receiver flask is slightly recessed, enabling capped flasks to be stacked for convenient storage.

NaviCyte

Harvard Apparatus www.harvardapparatus.com

Vertical diffusion/Ussing chamber systems

The Harvard/NaviCyte vertical diffusion chamber system is used for transport studies on both filter grown cell monolayers and surgically excised tissue sections. The vertical diffusion chamber system consists of six diffusion chambers, a 12-channel gas manifold and a heat block. Six general types of vertical chambers are available — one for use with cells grown on Snapwell membrane supports and the others for use with tissue.

ImmobaSil

Ashby Scientific www.ashby-scientific.co.uk

A cell support matrix for high productivity

ImmobaSil is a porous microcarrier support matrix for the immobilization and growth of both adherent and suspension cells. The porous structure provides a large surface area, and use of a gas-permeable silicon matrix enhances oxygen transfer, extending the lifespan of long-term cultures. ImmobaSil is claimed to be biocompatible with even the most delicate mammalian cell lines.

These notes are compiled in the Nature office from information provided by the manufacturers.

ADVERTISEMENTS

CUSTOMISED
monoclonal antibodies with
GUARANTEE

BioGenes
Gesellschaft für Biopolymere mbH

Internet www.biogenes.de
Berlin *Germany
Phone +49 (0)30-65 76 23 96

**DIVERSE
SMALL
MOLECULES
READY FOR
SCREENING**

**10K-50K
Compounds
in Plates**

- Next Day Delivery*
- Competitively Priced

*(Some Restrictions Apply)

Chembridge Corporation • 800.964.6143
858.451.7400 • 858.451.7401 FAX • E-mail: chem@chembridge.com

www.chembridge.com

Proteasome inhibition?

~ epoxomicin ~
~ PR39 ~ lactacystin ~
~ clasto-lactacystin ~ β ~
~ lactone ~ P131 ~
~ MG-132 ~ MG-262 ~

www.proteasome.com

AFFINITY Research Products Ltd.
Mamhead Castle, Exeter, EX6 8HD, U.K.
T. +44(0) 1626 891010 ~ F +44(0) 1626 891090
~ Email info@proteasome.com ~

BACs • YACs • PACs cDNAs

We provide clones and colony membranes for a number of human, mouse and rat libraries as well as a custom screening service for the libraries.

ResGen™
Invitrogen Corporation

U.S. or Canada 800-533-4363
FAX 256-536-9016

Check out our web site at
www.resgen.com

nature insight

Materials for clean energy



Cover illustration

Depleting supplies of fossil fuel and increasing future energy demands will require development of alternative 'clean' energy technologies, such as the solar panels shown in the background. (Images courtesy of M. Xeridat/I. Sullivan/Getty Images.)

Increasing awareness of environmental factors and limited energy resources have led to a profound evolution in the way we view the generation and supply of energy. Although fossil and nuclear sources will remain the most important energy provider for many more years, flexible technological solutions that involve alternative means of energy supply and storage need to be developed urgently.

The search for cleaner, cheaper, smaller and more efficient energy technologies has been driven by recent developments in materials science and engineering. The aim of this collection of reviews is therefore to focus on what materials-based solutions can offer and to show how the rational design and improvement of chemical and physical properties of these materials can lead to energy alternatives that can compete with existing technologies.

The most pronounced breakthroughs are currently taking place for technologies using renewable energy sources, such as fuel cells and solar cells. At the same time, the use of these technologies requires reliable and effective ways of storing energy, and exciting developments are occurring in the fields of hydrogen storage, rechargeable batteries and high-temperature superconductivity.

Exploring all the options under consideration would be impossible in such a collection. But we hope that these articles provide a flavour of the many scientific and technological challenges and future opportunities offered by alternative energy sources, as well as illustrating the multidisciplinary nature of the endeavour. Much work remains to be done before they start competing with conventional sources, but if we want continued prosperity and energy security while enjoying a safer and cleaner environment, we must face up to our responsibilities and ensure that these technologies start affecting our everyday life sooner rather than later.

We are pleased to acknowledge the financial support of the Materials Research Society in producing this Insight. As always, *Nature* has sole responsibility for editorial content and peer review.

Vincent Dusastre Senior Editor

overview:

332 Alternative energy technologies

*M. S. Dresselhaus
& I. L. Thomas*

review articles:

338 Photoelectrochemical cells

M. Grätzel

345 Materials for fuel-cell technologies

*B. C. H. Steele
& A. Heinzel*

353 Hydrogen-storage materials for mobile applications

*L. Schlapbach
& A. Züttel*

359 Issues and challenges facing rechargeable lithium batteries

*J.-M. Tarascon
& M. Armand*

368 High- T_c superconducting materials for electric power applications

*D. Larbalestier,
A. Gurevich,
D. M. Feldmann
& A. Polyanskii*

Editor, *Nature*: Philip Campbell
Insight Editors: Vincent Dusastre
Karl Ziemelis

Editorial Assistant: Simon Gibson
Production Editor: Simon Gribbin
Art Director: Majo Xeridat

Diagrams: Ann Thomson
Production Manager: Yvonne Strong
Sponsorship: Emma Jones

Alternative energy technologies

M. S. Dresselhaus* & I. L. Thomas†

*Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, Massachusetts 02139, USA

†Office of Basic Energy Sciences, U.S. Department of Energy, 19901 Germantown Road, Germantown, Maryland 20874-1290, USA

Fossil fuels currently supply most of the world's energy needs, and however unacceptable their long-term consequences, the supplies are likely to remain adequate for the next few generations. Scientists and policy makers must make use of this period of grace to assess alternative sources of energy and determine what is scientifically possible, environmentally acceptable and technologically promising.

Modern lifestyles demand a steady, reliable supply of energy: it lies at the heart of our mobility, our prosperity and our daily comfort. But we should not take this energy security for granted. Energy sources can be divided into three broad categories. The first derives from chemical or photophysical energy that relies on oxidizing some reduced substance, usually a hydrocarbon, or absorbing sunlight to generate either heat or electricity. The energy involved is that of a chemical bond or fractions of an electron volt (eV). The second involves nuclear reactions that release energy either by splitting heavy nuclei or by fusing light nuclei. The energy involved in nuclear reactions is in the region of 10^6 electron volts (MeV) per nuclear reaction. The third is thermomechanical in the form of wind, water, or geological sources of steam or hot water. The energy involved is in the milli-electron-volt (meV) region from, for example, water falling several tens of metres. The current usage of these various energy sources is depicted in Fig. 1.

Each energy source has some undesirable characteristics. Any process using fossil fuels produces carbon dioxide,

and perhaps also other contaminants, such as nitrogen oxides, sulphur oxides and ash. Nuclear plants produce radioactive fission products. Hydroelectric plants require dams and large lakes. Solar energy and wind energy require large areas and are limited geographically. Geothermal sources are limited to very few locations. Schemes using small temperature gradients in the earth or oceans have low thermal efficiencies, and hence require very large heat-exchanger areas.

At present most of the world's energy supply comes from fossil and nuclear sources (see Fig. 1). And although mankind is increasingly having to face the issues of resource limitation and environmental pollution, these sources will continue to be important in providing energy worldwide for the next few generations. Below, accordingly, we look briefly at the prospects for these energy sources, emphasizing the importance of addressing the issue of CO₂ sequestration now, and keeping the nuclear energy option open. But to meet increasing global demands for energy and to allow for the depletion of fossil fuel supplies in the coming years, alternative 'clean' energy sources, which do not depend on fossil fuels and which have a tolerable

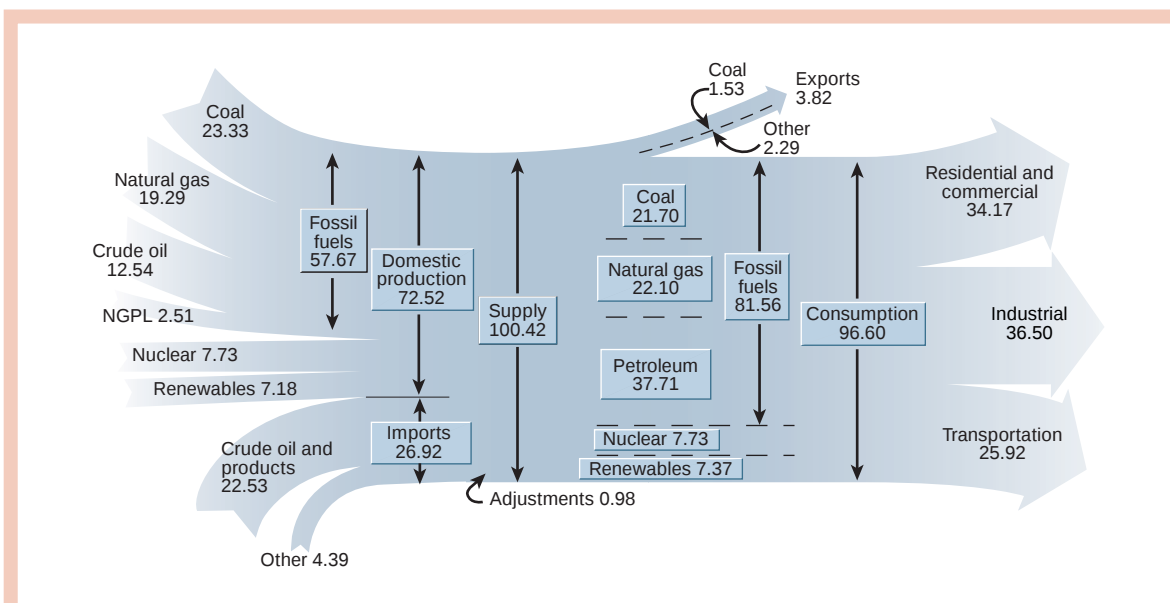


Figure 1 Energy flow diagram for the United States for 1999, in quads (1 quad = 10^{15} British thermal units = 2.9×10^{11} kWh). The average energy consumption in the United States is 0.4×10^{-6} quads per person per year, and the US population is about 5% of that on planet Earth. Energy consumption is large compared with food consumption (1.2×10^4 kJ per

day per person, which translates to only 0.4×10^{-8} quads per person per year). Some corresponding numbers for world energy consumption for 1999, in quads, are: petroleum 149.7; natural gas 87.3; coal 84.9; nuclear 25.2; hydro, geothermal, solar, wind and other renewables 29.9; total world energy production is 377.1 quads (ref. 8).



Figure 2 Dynamometer testing on the Honda Insight, a hybrid electric vehicle that has recently entered the market place.

environmental impact, must be developed. The five following articles^{1–5} review the present status of several such sources. Renewable means of producing and storing electricity are expected to be increasingly important in the future and, given the recent strides in condensed-matter physics and materials technology, could compete with existing technologies. For technologies using renewable energy sources, remarkable steps forward are taking place in the fields of fuel cells and solar cells. To use renewable sources effectively, reliable ways of storing energy are needed: exciting developments are being made in hydrogen storage, rechargeable batteries and high-temperature superconductivity. In this overview, we comment on the challenges and opportunities offered by the various alternative energy sources. Ultimately, the energy security of future generations will not only depend on reaching acceptable scientific and technological solutions, but will also require international cooperation on science policies to ensure continued prosperity and the safety of our environment.

Fossil energy

Although supplies are finite, currently there is no shortage of fossil fuels. World reserves of oil are about 1.6×10^{14} l (1×10^{12} barrels). World consumption is about 1.2×10^{10} l a day. World reserves of natural gas are about 1.4×10^{14} m³; gross production of gas is about 2.4×10^{12} m³ per year. World coal reserves are about 9.1×10^{11} tonnes; annual consumption is about 4.5×10^9 tonnes per year. The problems with fossil fuels for the near term are distribution and recovery, and these problems differ in detail for each of the fuel types. Much of the oil in a reservoir is not recovered because of surface tension. After primary recovery and water injection, most of the oil is left in pores as oil drops that are larger than the connecting necks of the porous formation. More could be recovered by injecting surfactants or pressurized CO₂ to reduce surface tension, but these approaches are expensive and technically challenging. Ultimately, the oil reservoir becomes a complicated research problem in the physics of multi-phase flow through porous bodies, solution chemistry, and critical phenomena in multi-component fluids. Underground coal mines have to leave large amounts of coal as structure to support the mine. Strip mining, which can remove essentially all of the coal, can have substantial environmental downsides. Gas is difficult to transport except through pipelines, and pipelines are not viable across oceans. It can be liquefied and transported more easily

as a cryogenic liquid, or converted to liquid fuels; however, better catalysts with greater selectivity are needed to make these processes cost-effective.

Finally, there are vast reserves of 'unconventional' fossil fuels, such as tar sands, oil shale and gas hydrates. Much research would need to be done to be able to extract energy from these economically and in an environmentally acceptable way. It is likely that the environmental cost of extracting and burning fossil fuels will mount too high, well before we run out of fossil fuels, and it is expected that it will take a long time to develop and deploy viable alternatives.

The main difficulty with fossil fuels is that they all produce CO₂ — natural gas the least, coal the most. It is possible to sequester the CO₂ — storing it as a gas or in solution underground or in the deep ocean, for example — but at some cost, both in money and in the energy expended in pumping the CO₂ from where it is produced to where it will be stored. The very idea of sequestration also raises questions. Would it prove stable in the long term? If there were a sudden release of sequestered CO₂, what would be the short-term consequences? Only increased research in geosciences can answer these questions. As fossil fuels will remain a dominant energy source for the near term, national policy to support research into CO₂ sequestration now is essential.

Electric cars represent one approach to reducing undesirable emissions from conventional internal combustion engines in heavily populated areas. As about two-thirds of the world's total electric power is generated from fossil fuels, electric cars ultimately consume mostly fossil fuels. But it is possible to produce electricity in areas of low population density, and electricity can be generated from sources, such as hydroelectric and nuclear energy, that do not produce greenhouse gases. The problem with electric cars is the storage of electric power. As yet, there is no battery that has the energy density of petrol, and battery recharging is far slower than filling a petrol tank. Also, the lifetimes of batteries need to be improved, and much lighter and non-toxic battery materials must be found to replace heavy metals like lead. An interesting concept that is now being tested is a hybrid vehicle with a small, very efficient petrol engine to recharge its batteries and to drive the vehicle after the electric motor has accelerated it. Braking also helps to charge the batteries; and because of the reserve electric motor, the petrol engine can be quickly restarted after every stop and need not idle, thereby resulting in high fuel efficiency (Fig. 2). Basic science is also playing an important role in,

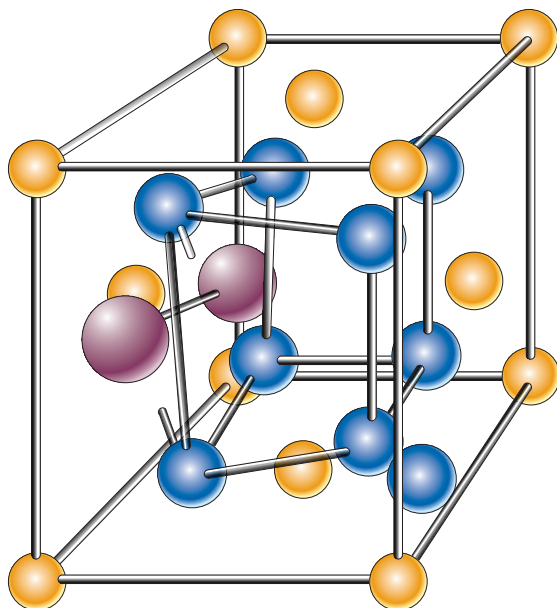


Figure 3 The structure of CeO_2 used in automobile catalytic converters as an oxygen source. Recent basic research⁹ using pulsed neutron scattering techniques combined with model calculations shows that the 'active oxygen' ion pairs for oxygenation are those found at interstices in the lattice – the octahedral sites (shown in purple). This result explains how to control oxygen in the catalytic converter so that NO_x can be reduced and CO can be oxidized in the same device. It also highlights how research at the most fundamental level is needed to address the problem of finding better catalytic materials.

for example, developing better catalytic converters for removing toxic species from automobile exhausts, while optimizing fuel efficiency. Even highly efficient engines produce oxides of nitrogen and carbon monoxide and will need catalytic converters (Fig. 3).

Nuclear energy

In the United States, nuclear energy from fission is now at a crossroads from which it will either set off on the path of recovery as a practical energy source, or slowly see each of its nuclear plants shut down as they reach the end of their useful life. The fission process does not emit any CO_2 , and from the standpoint of global warming, nuclear energy provides an ideal source. But the public has very real fears of nuclear energy that must be respected, understood and alleviated if this energy source is to remain viable (Fig. 1). Because of public sentiment, few nuclear power plants have been built worldwide in recent years. Reactors have to be made inherently safe and the problem of nuclear waste disposal must be solved. The scientific issues here are mostly in understanding long-term effects in the waste itself and in the repository, such as salt formation or stable rock formations. In the United States, the proposed site in Yucca Mountain is based on a volcanic rock called tuff. We need much better models to predict the stability of materials, such as glass, being irradiated for millennia and to predict the behaviour of the repositories under various situations that might arise (for example, a repository may flood or its humidity may change) until the radioactive elements decay to background levels. To extend reactor lifetimes, it is also important to understand the effect of radiation on the materials within operating reactors. Radiation continuously alters the structure of the material by displacing atoms and creating vacancies, often making a ductile material brittle. Nonetheless, science policy must be aligned with keeping the nuclear energy option open to address uncertain future energy needs.

Fusion is another nuclear process that, in principle, could provide essentially limitless power. The Sun demonstrates that fusion works, but the engineering problem of confining ions at temperatures high enough to overcome the Coulomb repulsion and achieve net energy output has not yet been solved on Earth. Although fusion is basically a physics problem, the engineering challenges are substantial. Fusion reactors using magnetic confinement technology as currently envisaged would use a mixture of deuterium and tritium as fuel. Tritium can be produced by bombarding lithium (either ^6Li or ^7Li) with a neutron to yield tritium (T) and an α -particle (^4He), and would have to be recovered and injected into the plasma. Many difficult materials problems arise because of radiation damage, especially to the materials exposed to the energetic plasma. Those materials will be bombarded by 14 MeV neutrons and by energetic neutral atoms that escape from the plasma. The first wall of the fusion reactor will have to withstand the high stresses that could occur if the plasma lost confinement. Again, much needs to be done to demonstrate the feasibility of fusion technology and to make it competitive with other options.

Alternative energy sources

As fossil fuel supplies are expected to be less available, more expensive and of increasing environmental concern in the coming century, increasing dependence on energy conservation and alternative energy sources is expected. The most obvious alternative energy source is the Sun. When the Sun is directly overhead and the sky is clear, radiation on a horizontal surface is about $1,000 \text{ W m}^{-2}$. To take the United States as an example, the total amount of solar energy falling on the continental 48 states is about 4.67×10^4 quads per year (a quad, derived from 'quadrillion', is $1.05 \times 10^{18} \text{ J}$ or $2.9 \times 10^{11} \text{ kWh}$), well in excess of the 98.6 quads that the United States consumes annually (Fig. 1). Currently about 0.08 quads is produced each year in the United States from solar thermal energy and from solar photovoltaics, which convert sunlight to electrical energy. Photovoltaics and solar cells have, in fact, provided reliable electrical power to space missions for many years. Sunlight can be used for heating, lighting and electricity generation, and it can be concentrated to provide steam to run turbines. In the summer and in the hotter areas of the United States, the availability of solar energy peaks when the demand for electricity peaks. The principal disadvantages of solar energy are that at present the conversion efficiency of sunlight to electric power is not high, and sunlight varies with time of day, weather conditions and season.

Photovoltaics and photoelectrochemical cells have not yet made a strong contribution to the energy supply because of their low conversion efficiencies and relatively high cost. Therefore, research has focused on developing materials with improved performance. Steady progress has been made in this quest. Grätzel, in this issue¹, provides a concise review of the historical background, and the physics and materials science issues that are involved in photoelectrochemical cells (Fig. 4). He also points to a number of newly emerging research opportunities that could substantially improve the performance and reduce the cost of photoelectrochemical cells or photovoltaics, thereby making it practical, for example, to use existing roofs as solar collectors. On average, over a large enough region, solar roofs tied to a grid could reduce peak daily electricity demand significantly. With the use of electronic band theory and modern computational capabilities, multi-junction photoelectric devices have been designed that use III–V compound semiconductors^{6,7} (Fig. 5) to achieve conversion efficiencies of over 30%. These impressive advances with inorganic solid-state junction devices are now being strongly challenged by new classes of materials, such as films containing nanocrystalline structures, quantum dots and nanostructured conducting polymers. These films typically contain nanoparticles with a size distribution showing a wide range of electronic band gaps (it is this gap that determines which wavelengths are absorbed), so that much of the solar spectrum can be captured by the solar cell. If such films can be made cheaply, with an

Figure 4 Photoelectrolysis occurring in a multi-junction cell, as a concentrated light source on the left is used to dissociate water to hydrogen and oxygen (which are observed as bubbles). The theoretical efficiency for tandem junction systems is 42% in converting the absorbed light into hydrogen production. Practical systems could achieve 18–24% efficiency; and low-cost multi-junction systems could achieve 5–12% efficiency^{10,11}.



optimized distribution of nanoparticle diameters, and can be properly aligned, one might have an ideal solar collector. This field is still very young, and is moving rapidly¹.

Closely related to solar cells are fuel cells, which convert fuel or chemical energy directly to electric energy. The main chemical reactions involve the oxidation of CO and H₂. As explained in the review article by Steele and Heinzel², fuel cells consist of an anode, cathode and electrolyte, just as in a typical battery. Although fuel cells have been operated reliably and efficiently in space missions for more than 40 years, they have not yet been widely used on the Earth, largely because of their cost. To make this technology commercially competitive, better anode, cathode and electrolyte materials and processes are needed². The various types of fuel cells now under development are critically reviewed by Steele and Heinzel², including polymeric electrolyte membranes and phosphoric acid fuel cells, both operating under ambient conditions, and the solid oxide and molten carbonate fuel cells, operating at higher temperatures. The efficiency of fuel cells is limited primarily by the diffusion of the fuel, oxygen and combustion products to and from the electrodes through the electrolyte. Phosphoric acid fuel cells are commercially available today and generate electricity at more than 40% efficiency. Polymer or proton-exchange-membrane fuel cells operating at about 90 °C are currently the best candidates for automotive use².

If we had better ways of storing the hydrogen, then hydrogen-based fuel cells would look much more promising². Most hydrogen is currently made by steam reforming of hydrocarbons. Hydrogen can also be made by heating coal to about 1,000 °C in the absence of oxygen, by partial burning of coal in the presence of steam, or by electrolysis. The reactions with coal make a mixture of H₂ with CO, CO₂ and other gases, which have to be separated from hydrogen. Generation of CO₂ can be avoided only if the electricity is generated from nuclear, hydroelectric, wind, solar or geothermal energy. The potential for photochemical production of hydrogen, using sunlight to dissociate water, is assessed in the review by Grätzel¹. Also under consideration are microbial pathways for the production of hydrogen (Fig. 6), a hint of the increasing role that the biosciences are expected to play in alternative energy technologies.

Hydrogen is attractive as a fuel because its oxidation product (water) is environmentally benign, it is lightweight and it is highly abundant. But storage is a problem, as Schlappbach and Züttel⁴

strongly emphasize. In their review article, the characteristics of hydrogen as a fuel are discussed, but the focus is on how we would store the hydrogen to be used as fuel for a fleet of automobiles, or for fuel cells. Recent interest in using carbon nanotubes as a storage medium, making use of their high surface area and light weight, is critically reviewed⁴, but the authors are not convinced that reversible hydrogen storage can be achieved in carbon nanotubes under

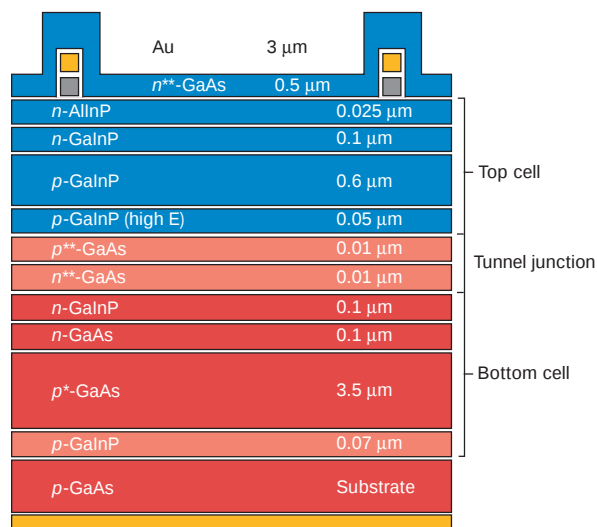


Figure 5 This multi-junction device uses a top cell of gallium indium phosphide and aluminium indium phosphide to absorb high-energy photons, transmitting lower-energy photons to the bottom cell where they are absorbed by the gallium arsenide and gallium indium phosphide junctions. A 'tunnel junction' is used in the middle of the device to aid the flow of electrons between the top and bottom cells. The different band gaps of the top and bottom junctions permit more efficient absorption of the solar spectrum. An anti-reflection coating and gold contacts are used above the top cell. This kind of solar cell has achieved over 30% efficiency under concentrated light^{6,7}.

Figure 6 Certain algae, whose chlorophyll give a green colour to the fluid, produce hydrogen in the presence of light. In engineered systems available today, 24% of the photosynthetic product is hydrogen rather than glucose. The photosynthesis process itself is about 1% efficient in converting light into chemical energy¹⁰.



reasonable operating conditions. They conclude that metal hydrides are the most promising materials for hydrogen storage, especially the lightweight metal alloy hydrides. Getting high hydrogen uptake is only part of the problem: much materials research will be needed to obtain controlled hydrogen release under practical conditions of temperature and pressure, and rapid hydrogen filling of the storage medium by recycling. Despite the interest and activity in hydrogen storage research, no breakthrough is yet in sight.

One alternative energy storage technology that has already taken off in a big way is the rechargeable lithium-ion battery. This technology has become a commercial reality through the efforts of the Sony Corporation and others, and is used widely today in portable computer and telecommunications devices, especially as these devices get smaller and more efficient. Demand and competition is driving the quest for higher storage capacity, longer operating times, faster recharging times, and other optimization of performance. Steady progress has been made by using improved materials for the anode, cathode and electrolyte, and the interfaces between them. Tarascon and Armand³ present an optimistic review of the technology, covering its historical background, its present status, and the challenges and opportunities now on the horizon. Three areas of opportunity, also common to most of the other alternative energy technologies, are compellingly presented: advances in nanostructured materials provide a chance to improve both anode and cathode performance; new *in situ* characterization techniques are helping to

identify materials problems in a format that permits rapid assessment of possible solutions; and advanced computer simulation modelling promises rapid progress in surveying new combinations of materials and geometries.

To get the best out of any alternate energy technology, improvements need to be made in energy storage and energy transmission. Superconductors are being seriously considered for both roles. For superconducting magnetic energy storage (SMES), electric energy is stored by circulating a current in a superconducting coil. Because there are no resistive losses, the current persists indefinitely. The efficiency of charging and discharging is very high because no energy conversion is involved, and for the same reason SMES can respond rapidly, with a response that is limited by the time required for conversion between d.c. and a.c. power. SMES can also be used for load levelling and for frequency and voltage control, which will probably be the first applications of this technology. Superconducting transmission lines could reduce resistive losses, but require energy for cryogenic cooling of the cables. It should be possible to use existing tunnels for these cables, thereby reducing cost while increasing the transmission capacity.

All large-scale applications of superconductors to energy storage and transmission will depend on achieving high critical current densities (J_{ct}) to be cost-competitive with other technologies. At present, niobium–titanium alloys are used for both storage and transmission at liquid helium temperatures. To use superconducting materials

with higher critical temperatures for these energy-related applications will require additional research and development. Larbalestier⁵, in his review, considers strategies for increasing J_c : in particular, by increasing the 'pinning' of magnetic flux lines to avoid energy losses from movement of magnetic vortices, while reducing losses owing to grain boundary and other defects. He also considers desirable attributes and criteria for selecting materials. Research is aimed at increasing the fraction of the cross-sectional area of the superconducting wire that is actually carrying the high current density. Both physics and materials science issues are being tackled. For all of these applications, high powers will be involved, so careful consideration must be given to the consequences of an accidental loss of coolant: an enormous amount of energy would be released the instant the superconductor became a normal conductor (10 MWh of stored energy is equivalent to 1.18 tonnes of TNT).

Research and science policy considerations

There is a great lack of public appreciation of the scientific and technological difficulties in supplying energy once fossil fuel sources become seriously depleted. The supply is adequate now and this gives us time to develop alternatives, but the scale of research in physics, chemistry, biology and engineering will need to be stepped up, because it will take sustained effort to solve the problem of long-term global energy security. To replace petrol or diesel oil with fuels of equivalent energy densities is going to take much greater understanding and control of chemical reactivity than we have now. To recycle CO₂, energy has to be converted to chemical bonds, and this conversion must be done very efficiently, for example as nature converts CO₂ and water to glucose using light. We need to understand energy and charge transport in molecular assemblies and across interfaces and membranes. We need to understand transition-state chemistry well enough to make selective and high-activity catalysts, essentially the equivalent of enzymes. And we need to understand how van der Waals forces, electrostatic forces and hydrogen bonding can be used to control molecular assembly and restrict reaction pathways.

Further great challenges lie in making the substantial improvements in materials that are needed to increase efficiency in the generation, conversion, transmission and use of energy. For example, the most effective way to increase fuel economy in automobiles is to lower their weight, and we are starting to see more vehicle components being made of lightweight, tough and strong composite materials. One reason for optimism is that physics, chemistry, biology and engineering research are now converging at the nanoscale. Many of the sub-cellular components that nature uses to make fuels and composite materials (wood, bone and shell, for instance) are at the nanoscale, including vital proteins, such as enzymes, photosynthetic units, molecular motors and cell membranes. For their part, physicists and chemists have developed many wonderful techniques to image, manipulate and interact with nanoscale objects in real time, and much progress

is being made in developing theoretical tools for studying assemblies of atoms on these scales.

In addition to advancing alternative energy technologies, research is also needed to manage fossil fuel resources better. Even though locating oil reservoirs using seismic techniques is highly developed, more effective oil extraction will require finer resolution and more accurate three-dimensional reconstruction of an image of an underground formation from the scattered wave. In exploiting any reservoir, mine or energy technology, we need to take environmental effects and waste disposal processes into consideration. Much more needs to be done to understand how various species flow and react in heterogeneous soils and formations. Complex chemical reactions in solution, reactions and ion exchange with soils, combined with flows through porous rocks in the presence of microorganisms, make this research very difficult.

These difficult scientific and technological problems in energy involve not only all of the sciences, but also science policy. Energy policy will differ among nations according to circumstance and national objectives, but many of the issues require international cooperation. All nations share the atmosphere of our planet, and all benefit from reduced emissions of pollutants. A healthy global economy benefits both producers and consumers of energy. International cooperation is especially important in basic research, and such cooperation should continue to be strongly encouraged. As energy plays such a vital role in society, it is important that policy, science and technology work together harmoniously; global energy security problems cannot be solved if the three components work in isolation. Policy determines what is acceptable, science shows what may be possible, and technology demonstrates what, within acceptable constraints, is practicable. □

1. Grätzel, M. Photoelectrochemical cells. *Nature* **414**, 338–344 (2001).
2. Steele, B. C. H. & Heinzel, A. Materials for fuel-cell technologies. *Nature* **414**, 345–352 (2001).
3. Tarascon, J.-M. & Armand, M. Issues and challenges facing rechargeable lithium batteries. *Nature* **414**, 359–367 (2001).
4. Schlapbach, L. & Züttel, A. Hydrogen-storage materials for mobile applications. *Nature* **414**, 353–358 (2001).
5. Larbalestier, D. High- T_c superconducting materials for electric power applications. *Nature* **414**, 368–377 (2001).
6. Bernard, J. E., Wei, S. H., Wood, D. M. & Zunger, A. Ordering-induced changes in the optical spectra of semiconductor alloys. *Appl. Phys. Lett.* **52**, 311–313 (1988).
7. Wei, S. H. & Zunger, A. Band-gap narrowing in ordered and disordered semiconductor alloys. *Appl. Phys. Lett.* **56**, 662–664 (1990).
8. Energy Information Administration Office of Energy Markets and End Use. *Annual Energy Review 1999* <<http://www.eia.doe.gov/aer>> (US Department of Energy, Washington DC, 2000); *International Energy Annual 1999* <<http://www.eia.doe.gov/aer>> (US Department of Energy, Washington DC, 2001).
9. Mamontov, E. & Egami, T. Structural defects in a nano-scale powder of CeO₂ studied by pulsed neutron diffraction. *J. Phys. Chem. Solids* **61**, 1345–1356 (2000).
10. Grätzel, M. in *CATTECH: the Magazine of Catalysis Sciences, Technology and Innovation* Issue 5 Vol. 3 No. 1, 4–17 (Baltzer Science Publishers, 1999).
11. International Energy Agency. *IEA Agreement on the Production and Utilization of Hydrogen*. 1999 Annual Report IEA/H2/AR-99 (ed. Elam, C. C.) p. 37 <<http://www.osti.gov/servlets/purl/774832-xO3oJD/webviewable/774832.pdf>> (1999).

Acknowledgements

We acknowledge fruitful discussions with many colleagues; G. Dresselhaus and P. Dehmer in particular have offered valuable suggestions.

Photoelectrochemical cells

Michael Grätzel

Institute of Photonics and Interfaces, Swiss Federal Institute of Technology, CH-1015, Lausanne, Switzerland (e-mail michael.gratzel@epfl.ch)

Until now, photovoltaics — the conversion of sunlight to electrical power — has been dominated by solid-state junction devices, often made of silicon. But this dominance is now being challenged by the emergence of a new generation of photovoltaic cells, based, for example, on nanocrystalline materials and conducting polymer films. These offer the prospect of cheap fabrication together with other attractive features, such as flexibility. The phenomenal recent progress in fabricating and characterizing nanocrystalline materials has opened up whole new vistas of opportunity. Contrary to expectation, some of the new devices have strikingly high conversion efficiencies, which compete with those of conventional devices. Here I look into the historical background, and present status and development prospects for this new generation of photoelectrochemical cells.

Ever since the French scientist Edmond Becquerel¹ discovered the photoelectric effect, researchers and engineers have been infatuated with the idea of converting light into electric power or chemical fuels. Their common dream is to capture the energy that is freely available from sunlight and turn it into the valuable and strategically important asset that is electric power, or use it to generate fuels such as hydrogen. Photovoltaics takes advantage of the fact that photons falling on a semiconductor can create

electron–hole pairs, and at a junction between two different materials, this effect can set up an electric potential difference across the interface. So far, the science of solar cells has been dominated by devices in which the junction is between inorganic solid-state materials, usually doped forms of crystalline or amorphous silicon, and profiting from the experience and material availability resulting from the semiconductor industry. Recently, we have seen more use of devices made from compound semiconductors — the III/V compounds for high-efficiency aerospace components and the copper–indium–sulphide/selenide materials for thin-film, low-cost terrestrial cells. But the dominance of the field by inorganic solid-state junction devices faces new challenges in the coming years. Increasingly, there is an awareness of the possible advantages of nanocrystalline and conducting polymer devices, for example, which are relatively cheap to fabricate (the expensive and energy-intensive high-temperature and high-vacuum processes needed for the traditional devices can be avoided), can be used on flexible substrates, and can be shaped or tinted to suit domestic devices or architectural or decorative applications. It is now even possible to depart completely from the classical solid-state junction device, by replacing the phase in contact with the semiconductor by an electrolyte (liquid, gel or organic solid), thereby forming a photoelectrochemical device.

The development of these new types of solar cells is promoted by increasing public awareness that the Earth's oil reserves could run out during this century. As the energy needs of the planet are likely to double within the next 50 years, the stage is set for a major energy shortage, unless renewable energy can cover the substantial deficit left by fossil fuels. Public concern has been heightened by the disastrous environmental pollution arising from all-too-frequent oil spills and the frightening climatic consequences of the greenhouse effect caused by fossil fuel combustion. Fortunately the supply of energy from the Sun to the Earth is gigantic: 3×10^{24} joules a year, or about 10,000 times more than the global population currently consumes. In other words, covering 0.1% of the Earth's surface with solar cells with an efficiency of 10% would satisfy our present needs. But to tap into this huge energy reservoir remains an enormous challenge.

Historical background

Becquerel's pioneering photoelectric experiments in 1839 were done with liquid not solid-state devices — a fact that is often ignored. His research, in which illumination of

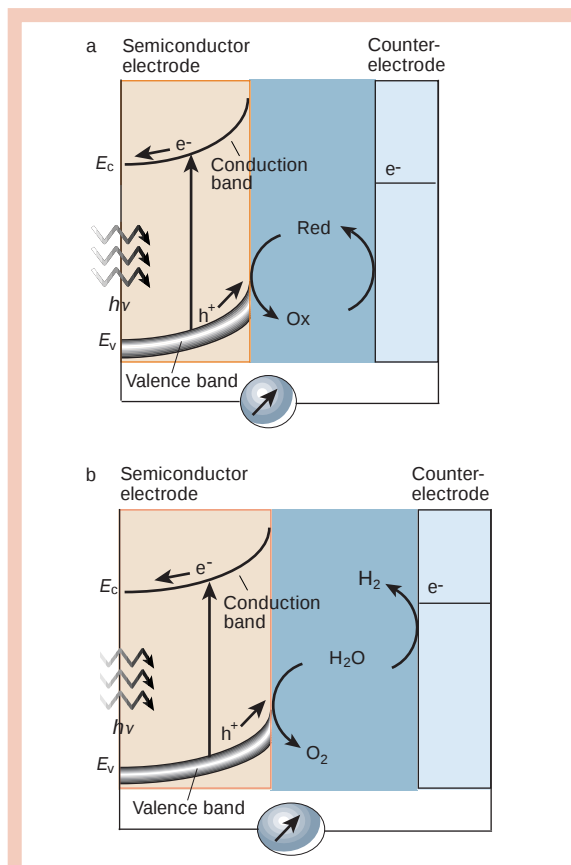
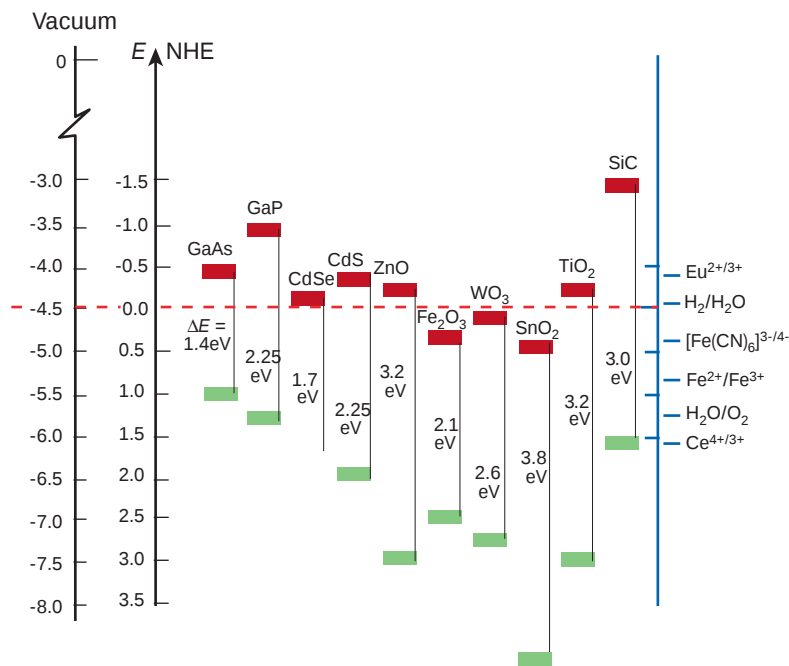


Figure 1 Principle of operation of photoelectrochemical cells based on n-type semiconductors. **a**, Regenerative-type cell producing electric current from sunlight; **b**, a cell that generates a chemical fuel, hydrogen, through the photo-cleavage of water.

Figure 2 Band positions of several semiconductors in contact with aqueous electrolyte at pH 1. The lower edge of the conduction band (red colour) and upper edge of the valence band (green colour) are presented along with the band gap in electron volts. The energy scale is indicated in electron volts using either the normal hydrogen electrode (NHE) or the vacuum level as a reference. Note that the ordinate presents internal and not free energy. The free energy of an electron–hole pair is smaller than the band gap energy due to the translational entropy of the electrons and holes in the conduction and valence band, respectively. On the right side the standard potentials of several redox couples are presented against the standard hydrogen electrode potential.



solutions containing a metal halide salt produced a current between two platinum electrodes immersed in the electrolyte, was motivated by photography. Daguerre had made the first photographic images in 1837, and Fox Talbot followed with the silver halide process in 1839. Although the art of formulating emulsions only became a science with the theoretical analysis of the process by Gurney and Mott² in 1938, there was constant empirical progress in the sensitivity of photographic films. Initially, films were particularly insensitive to mid-spectrum and red light. This is now recognized as being due to the semiconductor nature of the silver halide grains: they have a band gap (a gap in the allowed electronic energy levels) which ranges from 2.7 to 3.2 electron volts (eV) and negligible light absorption at wavelengths longer than 460 nm. Vogel's discovery³ in 1883 that silver halide emulsions could be sensitized by adding a dye extended the photosensitivity to longer wavelengths. Four years later, the concept of dye enhancement was carried over by Moser⁴ from photography to photoelectrochemical cells using the dye erythrosine on silver halide electrodes. This parallel between photography and photoelectrochemistry comes as a surprise to many chemists⁵. That the same dyes were particularly effective for both processes was recognized by Namba and Hishiki at the 1964 International Conference on Photosensitization of Solids in Chicago, which was a seminal event in the history of sensitization⁶. It was also recognized that the dye should be adsorbed on the semiconductor electrodes in a closely packed monolayer for maximum efficiency⁷. At this stage it was still debated whether sensitization occurred by transfer of electrons or of energy from the dye to the semiconductor⁸. Subsequent studies, notably by Hauffe⁹, Tributsch and Gerischer¹⁰, showed electron transfer to be the prevalent mechanism both for photographic and for photoelectrochemical sensitization processes.

Photosynthetic and regenerative cells

The foundation of modern photoelectrochemistry, marking its change from a mere support of photography to a thriving research direction on its own, was laid down by the work of Brattain and Garret¹¹ and subsequently Gerischer¹² who undertook the first detailed electrochemical and photoelectrochemical studies of the semiconductor–electrolyte interface. Research on photoelectrochemical cells went through a

frantic period after the oil crisis in 1973, which stimulated a worldwide quest for alternative energy sources. Within a few years well over a thousand publications appeared (see ref. 13 for a list).

Investigations focused on two types of cells whose principle of operation is shown in Fig. 1. The first type is the regenerative cell, which converts light to electric power leaving no net chemical change behind. Photons of energy exceeding that of the band gap generate electron–hole pairs, which are separated by the electric field present in the space-charge layer (see Box 1 for an explanation). The negative charge carriers move through the bulk of the semiconductor to the current collector and the external circuit. The positive holes are driven to the surface where they are scavenged by the reduced form of the redox relay molecule (R), oxidizing it: $h^+ + R \rightarrow O$. The oxidized form O is reduced back to R by the electrons that re-enter the cell from the external circuit. Much of the work on regenerative cells has focused on electron-doped (*n*-type) II/VI or III/V semiconductors using electrolytes based on sulphide/polysulphide, vanadium(II)/vanadium(III) or I₂/I⁻ redox couples. Conversion efficiencies of up to 19.6% have been reported for multijunction regenerative cells¹⁴.

The second type, photosynthetic cells, operate on a similar principle except that there are two redox systems: one reacting with the holes at the surface of the semiconductor electrode and the second reacting with the electrons entering the counter-electrode. In the example shown, water is oxidized to oxygen at the semiconductor photoanode and reduced to hydrogen at the cathode. The overall reaction is the cleavage of water by sunlight. Titanium dioxide has been the favoured semiconductor for these studies, following its use by Fujishima and Honda for water photolysis¹⁵. Unfortunately, because of its large band gap (3–3.2 eV, as shown in Fig. 2), TiO₂ absorbs only the ultraviolet part of the solar emission and so has low conversion efficiencies. Numerous attempts to shift the spectral response of TiO₂ into the visible, or to develop alternative oxides affording water cleavage by visible light, have so far failed.

In view of these prolonged efforts, disillusionment has grown about the prospects of photoelectrochemical cells being able to give rise to competitive photovoltaic devices, as those semiconductors with band gaps narrow enough for efficient absorption of visible light

Box 1

The semiconductor–electrolyte interface

When a semiconductor is placed in contact with an electrolyte, electric current initially flows across the junction until electronic equilibrium is reached, where the Fermi energy of the electrons in the solid (E_F) is equal to the redox potential of the electrolyte (E_{redox}), as shown in the figure. The transfer of electric charge produces a region on each side of the junction where the charge distribution differs from the bulk material, and this is known as the space-charge layer. On the electrolyte side, this corresponds to the familiar electrolytic double layer, that is, the compact (Helmholtz) layer followed by the diffuse (Gouy–Chapman) layer. On the semiconductor side of the junction the nature of the band bending depends on the position of the Fermi level in the solid. If the Fermi level of the electrode is equal to the flat band potential, there is no excess charge on either side of the junction and the bands are flat. If electrons accumulate at the semiconductor side one obtains an accumulation layer. If, however, they deplete from the solid into the solution, a depletion layer is formed, leaving behind a positive excess charge formed by immobile ionized donor states. Finally, electron depletion can go so far that their concentration at the interface falls below the intrinsic level. As a consequence, the semiconductor is *p*-type at the surface and *n*-type in the bulk, corresponding to an inversion layer. The illustration in the figure refers to *n*-type materials where electrons are the mobile charge carriers. For *p*-type semiconductors, analogous considerations apply. Positive holes are the mobile charge carriers and the immobile negatively charged states of the acceptor dopant form the excess space charge within the depletion layer.

The flat band potential is a very useful quantity in photoelectrochemistry as it facilitates location of the energetic position of the valence and conduction band edge of a given semiconductor material. It is obtained by measuring the capacity of the semiconductor–electrolyte junction. The semiconductor is subjected to reverse bias — that is, a voltage is applied to increase the potential step across the junction — and the differential capacity is determined as a function of the applied potential, V . The space-charge capacity of the semiconductor (C_{sc}) is in series with that of the Helmholtz layer (C_{H}) present at the electrolyte side of the interface. In the depletion regime the condition $C_{\text{H}} > C_{\text{sc}}$ applies, so the measured capacity is that of the space-charge layer. This depends on the applied bias voltage according to the Mott–Schottky equation: $1/(C_{\text{sc}})^2 = 2(\Delta\phi_{\text{sc}}RT/F)/(\epsilon_0\epsilon^1N)$, where $\Delta\phi_{\text{sc}} = V - V_{\text{fb}}$ is the voltage drop in the space-charge layer, R is the gas constant, F the Faraday number, ϵ the dielectric constant of the semiconductor, ϵ_0 the permittivity of vacuum and 1N the ionized donor dopant concentration. A plot of the square of the reciprocal capacity against the applied voltage gives a straight line and this is extrapolated to $1/(C_{\text{sc}})^2 = 0$ to derive the flat band potential V_{fb} .

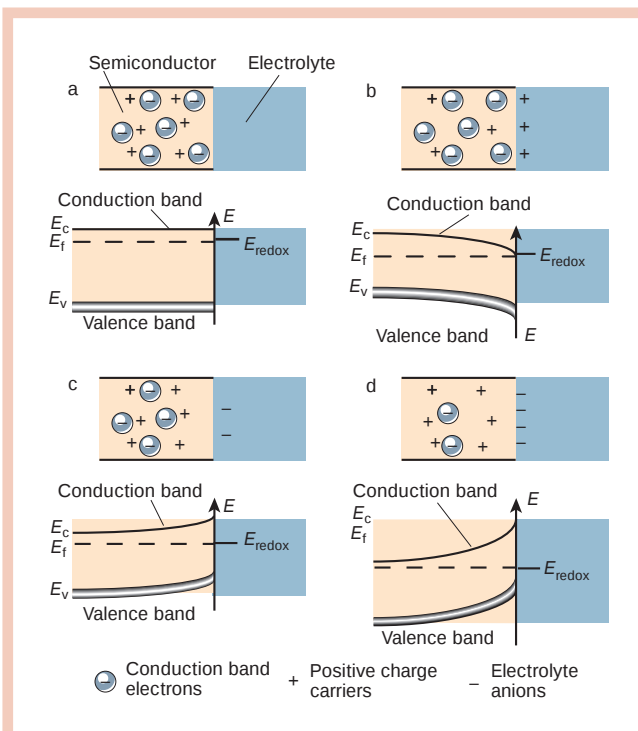
are unstable against photocorrosion. The width of the band gap is a measure of the chemical bond strength. Semiconductors stable under illumination, typically oxides of metals such as titanium or niobium, therefore have a wide band gap, an absorption edge towards the ultra-violet and consequently an insensitivity to the visible spectrum.

The resolution of this dilemma came in the separation of the optical absorption and charge-generating functions, using an electron transfer sensitizer absorbing in the visible to inject charge carriers across the semiconductor–electrolyte junction into a substrate with a wide band gap, and therefore stable. Figure 3 shows the operational principle of such a device.

Nanocrystalline junctions and interpenetrating networks

The need for dye-sensitized solar cells to absorb far more of the incident light was the driving force for the development of

Flat band potentials have been determined for a large number of materials⁴⁹ and some representative examples are shown in Fig. 2. Apart from the type of semiconductor they depend on the nature and composition of the electrolyte. In aqueous solution the flat band potentials of most oxide semiconductors shifts by 0.059 V when the pH is changed by one unit. This is a consequence of the fact that protons are potential-determining ions for these solids.



Box 1 Figure Schematic showing the electronic energy levels at the interface between an *n*-type semiconductor and an electrolyte containing a redox couple. The four cases indicated are: **a**, flat band potential, where no space-charge layer exists in the semiconductor; **b**, accumulation layer, where excess electrons have been injected into the solid producing a downward bending of the conduction and valence band towards the interface; **c**, depletion layer, where electrons have moved from the semiconductor to the electrolyte, producing an upward bending of the bands; and **d**, inversion layer where the electrons have been depleted below their intrinsic level, enhancing the upward band bending and rendering the semiconductor *p*-type at the surface.

mesoscopic semiconductor materials¹⁶ — minutely structured materials with an enormous internal surface area — which have attracted great attention during recent years. Mesoporous oxide films are made up of arrays of tiny crystals measuring a few nanometres across. Oxides such as TiO₂, ZnO, SnO₂ and Nb₂O₅, or chalcogenides such as CdSe, are the preferred compounds. These are interconnected to allow electronic conduction to take place. Between the particles are mesoscopic pores filled with a semiconducting or a conducting medium, such as a *p*-type semiconductor, a polymer, a hole transmitter or an electrolyte. The net result is a junction of extremely large contact area between two interpenetrating, individually continuous networks. Particularly intriguing is the ease with which charge carriers percolate across the mesoscopic particle network, making the huge internal surface area electronically addressable. Charge transport in such mesoporous

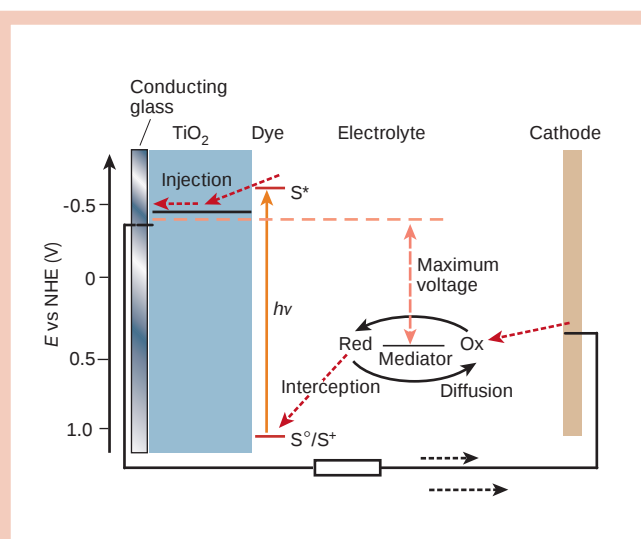


Figure 3 Schematic of operation of the dye-sensitized electrochemical photovoltaic cell. The photoanode, made of a mesoporous dye-sensitized semiconductor, receives electrons from the photo-excited dye which is thereby oxidized, and which in turn oxidizes the mediator, a redox species dissolved in the electrolyte. The mediator is regenerated by reduction at the cathode by the electrons circulated through the external circuit. Figure courtesy of P. Bonhôte/EPFL-LPI.

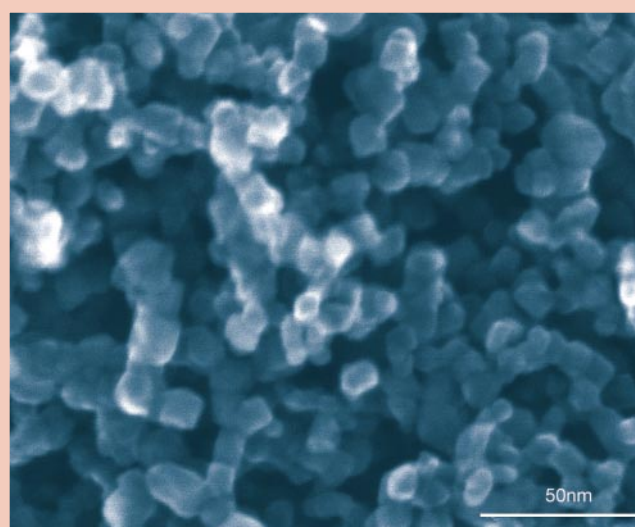


Figure 4 Scanning electron micrograph of the surface of a mesoporous anatase film prepared from a hydrothermally processed TiO_2 colloid. The exposed surface planes have mainly {101} orientation.

systems is under intense investigation today^{17,18} and is best described by a random walk model¹⁹.

The semiconductor structure, typically 10 μm thick and with a porosity of 50%, has a surface area available for dye chemisorption over a thousand times that of a flat, unstructured electrode of the same size. If the dye is chemisorbed as a monomolecular layer, enough can be retained on a given area of electrode to provide absorption of essentially all the incident light. Figure 4 shows an electron micrograph of a nanocrystalline TiO_2 film with a grain size in the range of 10–80 nm. (For details of the hydrothermal deposition procedure, starting with hydrolysis of a titanium isopropoxide precursor and terminating with screen printing and firing of the semiconductor layer on a conductive transparent substrate, see ref. 20.)

The nanostructuring of the semiconductor introduces profound changes in its photoelectrochemical properties. Of great importance is the fact that a depletion layer (see Box 1) cannot be formed in the solid — the particles are simply too small. The voltage drop within the nanocrystals remains small under reverse bias, typically a few millivolts. As a consequence there is no significant local electric field present to assist in the separation of photogenerated electron–hole pairs. The photoresponse of the electrode is determined by the rate of reaction of the positive and negative charge carriers with the redox couple present in the electrolyte. If the transfer of electrons to the electrolyte is faster than that of holes, then a cathodic photocurrent will flow, like in a *p*-type semiconductor/liquid junction. In contrast, if hole transfer to the electrolyte is faster, then anodic photocurrent will flow, as in *n*-type semiconductor photoelectrochemical cells. Striking confirmation of the importance of these kinetic effects came with the demonstration²¹ that the same nanocrystalline film could show alternatively *n*- or *p*-type behaviour, depending on the nature of the hole or electron scavenger present in the electrolyte phase. This came as a great surprise to a field where the traditional thinking was to link the photoresponse to formation of a charge-depletion layer at the semiconductor–electrolyte interface.

What, then, is the true origin of the photovoltage in dye-sensitized solar cells? In the conventional picture, the photovoltage of photoelectrochemical cells does not exceed the potential drop in the space-charge layer (Box 1 Figure). But nanocrystalline cells can

develop photovoltages close to 1 V even though the junction potential is in the millivolt range. It has been suggested that a built-in potential difference at the back contact of the nanocrystalline film with the conducting glass is responsible for the observed photovoltage²². Other evidence²³ suggests that under illumination, electron injection from the sensitizer increases the electron concentration in the nanocrystalline electrode, raising the Fermi level of the oxide and thus shifting its potential²⁴. From recent electrical impedance studies²⁵, it seems that both changes — the potential drop across the back contact and the Fermi level shift of the TiO_2 nanoparticles — contribute to the photovoltage of dye-sensitized solar cells.

Accumulation layers (part b of Box 1 Figure) can be produced in the nanocrystals under forward bias when majority carriers are injected, rendering the film highly conductive. Under reverse bias the carriers are withdrawn, turning it into an insulator. Thus, by changing the applied potential, the film can be switched back and forth from a conducting to an insulating state. Space-charge limitation of the current (arising from limitation of the density of charge carriers because they are repelled by each other's electric field) is not observed as the injected majority carriers are efficiently screened by the electrolyte present in the pores surrounding the nanoparticles. The factors controlling the rate of charge carrier percolation across the nanocrystalline film are under intense scrutiny¹⁷. A technique known as intensity-modulated impedance spectroscopy has proved to be an elegant and powerful tool^{25,26} for addressing these and other important questions related to the characteristic time constants for charge carrier transport and reaction dynamics.

An interesting feature specific to nanocrystalline electrodes is the appearance of quantum confinement effects. These appear when the films are made up of small quantum dots, such as 8-nm-sized CdTe particles²⁷. Such layers have a larger band gap than the bulk material, the band edge position being shifted with respect to the positions indicated in Fig. 2 for macroscopic materials. The conduction band redox potential is lowered and that of the valence band is increased. As a consequence, electrons and holes can perform reduction and oxidation reactions that cannot proceed on bulk semiconductors.

The astounding photoelectrochemical performance of nanocrystalline semiconductor junctions is illustrated in Fig. 5 where we

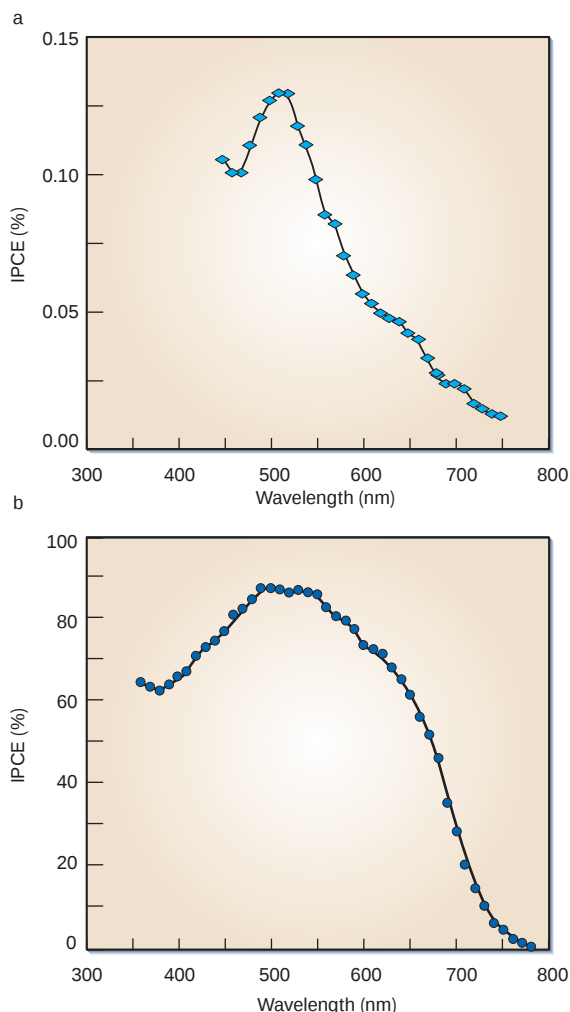


Figure 5 The nanocrystalline effect in dye-sensitized solar cells. In both cases, TiO₂ electrodes are sensitized by the surface-anchored ruthenium complex *cis*-RuL₂(SCN)₂. The incident-photon-to-current conversion efficiency is plotted as a function of the excitation wavelength. **a**, Single-crystal anatase cut in the (101) plane. **b**, Nanocrystalline anatase film. The electrolyte consisted of a solution of 0.3M LiI and 0.03M I₂ in acetonitrile.

compare the photoresponse of an electrode made of single-crystal anatase, one of the crystal forms of TiO₂, with that of a mesoporous TiO₂ film. Both electrodes are sensitized by the ruthenium complex *cis*-RuL₂(SCN)₂ (L is 2,2'-bipyridyl-4,4'-dicarboxylate), which is adsorbed as a monomolecular film at the titania surface. The incident-photon-to-current conversion efficiency (IPCE) is plotted as a function of wavelength. The IPCE value obtained with the single-crystal electrode is only 0.13% near 530 nm, where the sensitizer has an absorption maximum, whereas it reaches 88% with the nanocrystalline electrode²⁸ — more than 600 times as great. The photocurrent in standard sunlight augments 10³–10⁴ times when passing from a single crystal to a nanocrystalline electrode (standard, or full, sunlight is defined as having a global intensity (*I*_s) of 1,000 W m⁻², air mass 1.5; air mass is the path length of the solar light relative to a vertical position of the Sun above the terrestrial absorber). This striking improvement is due largely to the far better light harvesting of the dye-sensitized nanocrystalline film as compared with a flat single-crystal electrode, but is also due, at least in part, to the mesoscopic film texture favouring photogeneration and collection of charge carriers²⁹.

The overall conversion efficiency of the dye-sensitized cell is determined by the photocurrent density measured at short circuit (*i*_{ph}), the open-circuit photo-voltage (*V*_{oc}), the fill factor of the cell (*ff*) and the intensity of the incident light (*I*_s)

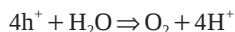
$$\eta_{\text{global}} = i_{\text{ph}} V_{\text{oc}} (ff/I_s)$$

Under full sunlight, short-circuit photocurrents ranging from 16 to 22 mA cm⁻² are reached with state-of-the-art ruthenium sensitizers, while *V*_{oc} is 0.7–0.8 V and the fill factor values are 0.65–0.75. A certified overall power conversion efficiency of 10.4% has been attained at the US National Renewable Energy Laboratory³⁰. Although this efficiency makes dye-sensitized cells fully competitive with the better amorphous silicon devices, an even more significant parameter is the dye lifetime achieved under working conditions. For credible system performance, a dye molecule must sustain at least 10⁸ redox cycles of photo-excitation, electron injection and regeneration, to give a device service life of 20 years. The use of solvents such as valeronitrile, or γ -butyrolactone, appropriately purified, in the electrolyte formulation provide a system able to pass the standard stability qualification tests for outdoor applications, including thermal stress for 1,000 h at 85 °C, and this has been verified independently³¹.

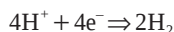
Tandem cells for water cleavage by visible light

The advent of nanocrystalline semiconductor systems has rekindled interest in tandem cells for water cleavage by visible light, which remains a highly prized goal of photoelectrochemical research. The 'brute force' approach to this goal is to use a set of four silicon photovoltaic cells connected in series to generate electricity that is subsequently passed into a commercial-type water electrolyser. Solar-to-chemical conversion efficiencies obtained are about 7%. Much higher efficiencies in the range of 12–20% have been reported for tandem cells based on III/V semiconductors^{14,32}, but these single-crystal materials cost too much for large-scale terrestrial applications.

A low-cost tandem device that achieves direct cleavage of water into hydrogen and oxygen by visible light was developed recently³³. This is based on two photosystems connected in series as shown in the electron flow diagram of Fig. 6. A thin film of nanocrystalline tungsten trioxide, WO₃ (ref. 34), or Fe₂O₃ (ref. 35) serves as the top electrode absorbing the blue part of the solar spectrum. The valence-band holes (h⁺) created by band-gap excitation of the film oxidize water to oxygen:



and the conduction-band electrons are fed into the second photosystem consisting of the dye-sensitized nanocrystalline TiO₂ cell discussed above. The latter is placed directly under the WO₃ film, capturing the green and red part of the solar spectrum that is transmitted through the top electrode. The photovoltage generated by the second photosystem enables hydrogen to be generated by the conduction-band electrons.



The overall reaction corresponds to the splitting of water by visible light. There is close analogy to the 'Z-scheme' (named for the shape of the flow diagram) that operates in photosynthesis. In green plants, there are also two photosystems connected in series, one that oxidizes water to oxygen and the other generating the compound NADPH used in fixation of carbon dioxide. As discussed above, the advantage of the tandem approach is that higher efficiencies can be reached than with single junction cells if the two photosystems absorb complementary parts of the solar spectrum. At present, the overall conversion efficiency from standard solar light to chemical

energy achieved with this device stands at 4.5%, and further improvements are being sought.

Dye-sensitized solid heterojunctions and ETA cells

Interest is growing in devices in which both the electron- and hole-carrying materials are solids, but are grown as interpenetrating networks forming a heterojunction of large contact area. From conventional wisdom one would have predicted that solar cells of this kind would work very poorly, if at all. The disordered character of the junction and the presence of the huge interface are features one tries to avoid in conventional photovoltaic cells, because the disruption of the crystal symmetry at the surface produces electronic states in the band gap of the semiconductor, enhancing the recombination of photogenerated carriers. The fact that molecular photovoltaic cells based on the sensitization of nanocrystalline TiO_2 were able to achieve overall conversion efficiencies from solar to electric power of over 10% encouraged work on solid-state analogues, that is, dye-sensitized heterojunctions. The first devices of this type used inorganic *p*-type semiconductors, for example CuI (ref. 36) or CuSCN (ref. 37), as hole conductors replacing the redox electrolyte. Respectable conversion efficiencies exceeding 1% have been reached with such cells. But the lack of photostability of the Cu(I) compounds and the difficulty of realizing a good contact between the two mesoscopic inorganic materials still present considerable practical challenges.

Organic charge-transport materials have advantages in this respect. An amorphous hole conductor can be introduced into the mesoporous TiO_2 film by a simple spin-coating process and readily adapts its form to the highly corrugated oxide surface. Cells based on a spirobisfluorene-connected arylamine hole transmitter³⁸, which fills the pores of a dye-sensitized nanocrystalline TiO_2 film, have reached a conversion efficiency of 2.56% at full sunlight³⁹. The high open-circuit voltage of these devices, exceeding 900 mV, is particularly noteworthy and promising for further substantial improvements in performance. In general, dye-sensitized heterojunction cells offer great flexibility because the light absorber and charge-transport material can be selected independently to obtain optimal solar energy harvesting and high photovoltaic output. The great advantage of such a configuration is that the charge carriers are generated by the dye precisely at the site of the junction where the electric field is greatest, enhancing charge separation.

Extremely thin absorber (ETA) solar cells are conceptually close to dye-sensitized heterojunctions. The molecular dye is replaced by an extremely thin (2–3 nm) layer of a small-band-gap semiconductor, such as CuInS_2 . A hole conductor such as CuSCN is placed on top of the absorber, producing a junction of the PIN type (*p*-type semiconductor/insulator/*n*-type semiconductor)⁴⁰. The structure has the advantage of enhanced light harvesting due to the surface enlargement and multiple scattering. Because photo-induced charge separation occurs on a length scale of a few nanometres, higher levels of defects and impurities can be tolerated than in flat thin-film devices, where the minority carriers are required to diffuse several microns. On the other hand, making PIN-junctions of such high contact area is difficult and this has hampered the performance of these cells. Their conversion efficiency so far has remained below 5%, which is less than one-third of the yield obtained with similar semiconductor materials in a flat junction configuration.

Soft junctions and organic solar cells

Organic materials have the advantage of being cheap and easy to process. They can be deposited on flexible substrates, bending where their inorganic competitors would crack. The choice of materials is practically unlimited, and specific parts of the solar spectrum can be selectively absorbed. Although organic cells are still considerably less efficient than single-crystal gallium arsenide or silicon, progress has been impressive over the past few years. In particular, solar cells based on interpenetrating polymer networks⁴¹, polymer/fullerene blends⁴²,

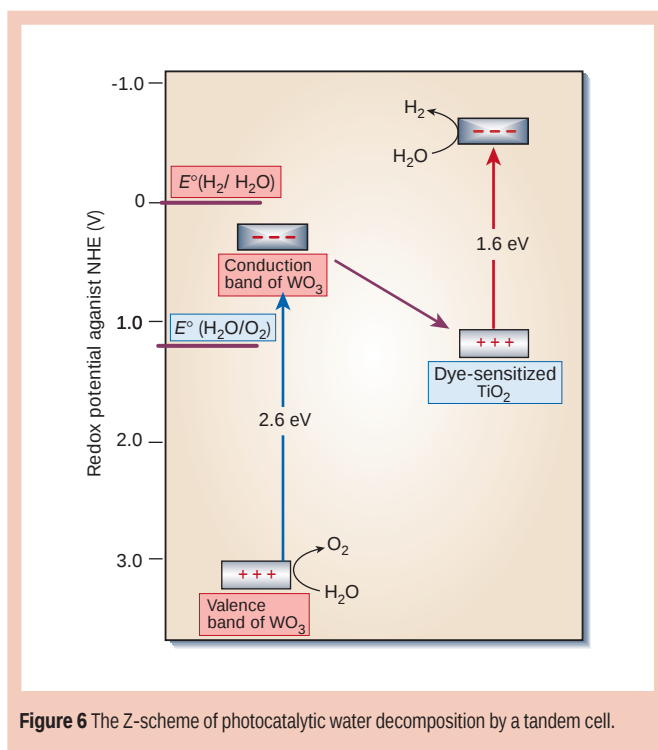


Figure 6 The Z-scheme of photocatalytic water decomposition by a tandem cell.

halogen-doped organic crystals⁴³ and the solid-state dye-sensitized devices mentioned above³⁸ have shown surprisingly high solar conversion efficiencies, currently reaching values of 2–3%.

Conducting polymers, for example poly-(phenylenevinylene) (PPV) derivatives or C_{60} particles, are attracting great interest as photovoltaic materials^{44,49}. Bulk donor–acceptor heterojunctions are formed simply by blending the two organic materials serving as electron donor (*p*-type conductor) and electron acceptor (*n*-type conductor). The advantage of these new structures over the flat-junction organic solar cells investigated earlier⁴⁵ is the interpenetration of the two materials that conduct positive and negative charge carriers, reducing the size of the individual phase domains to the nanometre range. This overcomes one of the problems of the first generation of organic photovoltaic cells: the unfavourable ratio of exciton diffusion length to optical absorption length. An exciton is a bound electron–hole pair produced by absorption of light; to be useful, this pair must reach the junction and there dissociate into two free charge carriers — but excitons typically diffuse only a few nanometres before recombining. Light is absorbed (and generates excitons) throughout the composite material. But in the composite, the distance the exciton has to travel before reaching the interface is at most a few nanometres, which is commensurate with its diffusion length. Hence photo-induced charge separation can occur very efficiently. A conversion efficiency from incident photons to current of over 50% has been achieved with a blend containing PPV and methanofullerene derivatives⁴⁶. The overall conversion efficiency from solar to electric power under full sunlight achieved with this cell was 2.5%. Although these results are impressive, the performance of the cell declined rapidly within hours of exposure to sunlight⁴⁷. In contrast, the output of dye-sensitized solar cells is remarkably stable even under light soaking for more than 10,000 h. Similar long-term stability will be required for large-scale application of polymer solar cells.

Summary

Photovoltaic devices based on interpenetrating mesoscopic networks have emerged as a credible alternative to conventional solar cells. Common to all these cells is an ultrafast initial charge separation step, occurring in femtoseconds, and a much slower back-reaction.

Table 1 Performance of photovoltaic and photoelectrochemical solar cells

Type of cell	Efficiency (%) ^a		Research and technology needs
	Cell	Module	
Crystalline silicon	24	10–15	Higher production yields, lowering of cost and energy content
Multicrystalline silicon	18	9–12	Lower manufacturing cost and complexity
Amorphous silicon	13	7	Lower production costs, increase production volume and stability
CuInSe ₂	19	12	Replace indium (too expensive and limited supply), replace CdS window layer, scale up production
Dye-sensitized nanostructured materials	10–11	7	Improve efficiency and high-temperature stability, scale up production
Bipolar AlGaAs/Si photoelectrochemical cells	19–20	—	Reduce materials cost, scale up
Organic solar cells	2–3	—	Improve stability and efficiency

^aEfficiency defined as conversion efficiency from solar to electrical power.

This allows the charge carriers to be collected as electric current before recombination takes place. Table 1 compares the performance of the new photoelectrochemical systems with conventional devices. Although still of lower efficiency, the nanostructured cells offer several advantages over their competitors. They can be produced more cheaply and at less of a cost in energy than silicon cells, for which 5 GJ have to be spent to make 1 m² of collector area. Unlike silicon, their efficiency increases with temperature, narrowing the efficiency gap under normal operating conditions. They usually have a bifacial configuration, allowing them to capture light from all angles. Transparent versions of different colour can readily be made that could serve as electric power-producing windows in buildings. These and other attractive features⁴⁸ justify the present excitement about these cells and should aid their entry into a tough market. □

- Bequerel, E. Recherches sur les effets de la radiation chimique de la lumière solaire, au moyen des courants électriques. *C.R. Acad. Sci.* **9**, 145–149 (1839).
- Gurney, R. W. & Mott, N. F. Theory of the photolysis of silver bromide and the photographic latent image. *Proc. R. Soc. Lond. A* **164**, 151–167 (1938).
- West, W. First hundred years of spectral sensitization. *Proc. Vogel Cent. Symp. Photogr. Sci. Eng.* **18**, 35–48 (1974).
- Moser, J. Notiz über die Verstärkung photoelectrischer Ströme durch optische Sensibilisierung. *Monatsh. Chem.* **8**, 373 (1887).
- Spitler, M. T. Dye photo-oxidation at semiconductor electrodes—a corollary to spectral sensitization in photography. *J. Chem. Educ.* **60**, 330–332 (1983).
- Namba, S. & Hishiki, Y. Color sensitization of zinc oxide with cyanide dyes. *J. Phys. Chem.* **69**, 774–779 (1965).
- Nelson, R. C. Minority carrier trapping and dye sensitization. *J. Phys. Chem.* **69**, 714–718 (1965).
- Boudon, J. Spectral sensitization of chemical effects in solids. *J. Phys. Chem.* **69**, 705–713 (1965).
- Gerischer, H. & Tributsch, H. Electrochemische Untersuchungen zur spectraleu sensibilisierung von ZnO-Einkristallen. *Ber. Bunsenges. Phys. Chem.* **72**, 437–445 (1968).
- Hauffe, K., Danzmann, H. J., Pusch, H., Range, J. & Volz, H. New Experiments on the sensitization of zinc oxide by means of the electrochemical cell technique. *J. Electrochem. Soc.* **117**, 993–999 (1970).
- Brattain, W. H. & Garrett, C. G. B. Experiments on the interface between germanium and an electrolyte. *Bell Syst. Tech. J.* **34**, 129–176 (1955).
- Gerischer, H. Electrochemical behavior of semiconductors under illumination. *J. Electrochem. Soc.* **113**, 1174–1182 (1966).
- Kalyanasundaram, K. Photoelectrochemical cell studies with semiconductor electrodes: a classified bibliography (1975–1983). *Solar Cells* **15**, 93–156 (1985).
- Licht, S. Multiple band gap semiconductor/electrolyte solar energy conversion. *J. Phys. Chem.* **105**, 6281–6294 (2001).
- Fujishima, A. & Honda, K. Electrochemical photolysis of water at a semiconductor electrode. *Nature* **238**, 37–38 (1972).
- O'Regan, B. & Grätzel, M. A low-cost, high efficiency solar cell based on dye-sensitized colloidal TiO₂ films. *Nature* **353**, 737–740 (1991).
- Hagfeldt, A. & Grätzel, M. Molecular photovoltaics. *Acc. Chem. Res.* **33**, 269–277 (2000).
- Hilgendorff, M., Spanhel, L., Rothenhäusler, Ch. & Müller, G. From ZnO colloids to nanocrystalline highly conductive films. *J. Electrochem. Soc.* **145**, 3632–3637 (1998).
- Nelson, J. Continuous time random walk model of electron transport in nanocrystalline TiO₂ electrodes. *Phys. Rev. B* **59**, 15374–15380 (1999).
- Barbé, Ch. J. et al. Nanocrystalline titanium oxide electrodes for photovoltaic applications. *J. Am. Ceram. Soc.* **80**, 3157–3171 (1997).
- Hodes, G., Howell, I. D. J. & Peter, L. M. Nanocrystalline photoelectrochemical cells: a new concept in photovoltaic cells. *J. Electrochem. Soc.* **139**, 3136–3140 (1992).
- Schwarzburg, K. & Willig, F. Origin of photovoltage in dye sensitized electrochemical solar cells. *J. Phys. Chem. B* **103**, 5743–5746 (1999).
- Pichot, F. & Gregg, B. A. The photovoltage-determining mechanism in dye-sensitized solar cells. *J. Phys. Chem. B* **104**, 6–10 (1999).
- Cahen, D., Hodes, G., Grätzel, M., Guillemoles, J. F. & Riess, I. Nature of photovoltaic action in dye-sensitized solar cells. *J. Phys. Chem. B* **104**, 2053–2059 (2000).
- Van de Lagemaat, J., Park, N.-G. & Frank, A. J. Influence of electrical potential distribution, charge transport, and recombination on the photopotential and photocurrent conversion efficiency of dye-sensitized nanocrystalline TiO₂ solar cells: a study by electrical impedance and optical modulation techniques. *J. Phys. Chem. B* **104**, 2044–2052 (2000).
- Dłoczik, L. et al. Dynamic response of dye-sensitized nanocrystalline solar cells: characterization by intensity-modulated photocurrent spectroscopy. *J. Phys. Chem. B* **101**, 10281–10289 (1997).
- Masatai, Y. & Hodes, G. Size quantization in electrodeposited CdTe nanocrystalline films. *J. Phys. Chem. B* **101**, 2685–2690 (1997).
- Nazeruddin, M. K. et al. Conversion of light to electricity by cis X2-Bis(2,2'-bipyridyl-4,4'-dicarboxylate) ruthenium(II) charge transfer sensitizers. *J. Am. Chem. Soc.* **115**, 6382–6390 (1993).
- Kavan, L., Grätzel, M., Gilbert, S. E., Klemenz, C. & Scheel, H. J. Electrochemical and photoelectrochemical investigations of single-crystal anatase. *J. Am. Chem. Soc.* **118**, 6716–6723 (1996).
- Nazeeruddin, M. K. et al. Engineering of efficient panchromatic sensitizers for nanocrystalline TiO₂-based solar cells. *J. Am. Chem. Soc.* **123**, 1613–1624 (2001).
- Hinsch, A. et al. in *Proc. 16th Eur. PV Solar Energy Conf., Glasgow, May 2000* (eds Scheel, H. et al.) 32 (James & James, London, 2000).
- Khaselev, O. & Turner, J. A. A monolithic photovoltaic-photoelectrochemical device for hydrogen production via water splitting. *Science* **280**, 425–427 (1998).
- Grätzel, M. The artificial leaf, bio-mimetic photocatalysis. *Cattech* **3**, 3–17 (1999).
- Santato, C., Ulmann, M. & Augustynski, J. Photoelectrochemical properties of nanostructured tungsten trioxide films. *J. Phys. Chem. B* **105**, 936–940 (2001).
- Khan, S. U. M. & Akikusa, J. Photoelectrochemical splitting of water at nanocrystalline n-Fe₂O₃ thin-film electrodes. *J. Phys. Chem. B* **103**, 7184–7189 (1999).
- Tennakone, K., Kumara, G. R. R. A., Kumarasinghe, A. R., Wijayantha, K. G. U. & Sirimanne, P. M. A dye-sensitized nano-porous solid-state photovoltaic cell. *Semicond. Sci. Technol.* **10**, 1689–1693 (1995).
- O'Regan, B. & Schwarz, D. T. Large enhancement in photocurrent efficiency caused by UV illumination of dye sensitized heterojunctions. *Chem. Mater.* **10**, 1501–1509 (1998).
- Bach, U. et al. Solid-state dye-sensitized mesoporous TiO₂ solar cells with high photon-to-electron conversion efficiencies. *Nature* **395**, 583–585 (1998).
- Krüger, J., Bach, U. & Grätzel, M. High efficiency solid-state photovoltaic device due to inhibition of interface charge recombination. *Appl. Phys. Lett.* **79**, 2085–2087 (2001).
- Kaiser, I. et al. The eta-solar cell with CuInS₂: a photovoltaic cell concept using an extremely thin absorber. *Sol. Energy Mater. Sol. Cells* **67**, 89–96 (2001).
- Halls, J. J. M., Pickler, K., Friend, R. H., Morati, S. C. & Holmes, A. B. Efficient photodiodes from interpenetrating polymer networks. *Nature* **376**, 498–500 (1995).
- Yu, G., Gao, J., Hummelen, J. C., Wudl, F. & Heeger, A. J. Polymer photovoltaic cells: enhanced efficiencies via a network of internal donor acceptor heterojunctions. *Science* **270**, 1789–1791 (1995).
- Schön, J. H., Kloc, Ch., Bucher, E. & Batlogg, B. Efficient organic photovoltaic diodes based on doped pentacene. *Nature* **403**, 408–410 (2000).
- Brabec, C. J. & Sariciftci, N. S. Polymeric photovoltaic devices. *Mater. Today* 3–8 (2000).
- Wöhrle, D. & Meissner D. Organic solar cells. *Adv. Mat.* **3**, 129–138 (1991).
- Shaheen, S. E. et al. 2.5% efficient organo-nic plastic solar cells. *Appl. Phys. Lett.* **78**, 841–843 (2001).
- Tuladhar, D. et al. Abstract, Int. Workshop Nanostruct. Photovoltaics, Dresden, Germany <<http://www.mpiikps-dresden.mpg.de>> (2001).
- Grätzel, M. Perspectives for dye-sensitized nanocrystalline solar cells. *Prog. Photovoltaic Res. Appl.* **8**, 171–185 (2000).
- Savenije, T. J., Warman, J. M. & Goossens, A. Visible light sensitization of titanium dioxide using a phenylene vinylene polymer. *Chem. Phys. Lett.* **278**, 148–153 (1998).

Acknowledgements

I thank the members of the Swiss Federal Institute of Technology (EPFL) electrochemical photovoltaics development team, some of whose work is referenced; the industrial organizations whose interest in this PV system has induced them to license the concept and thereby support the research; EPFL; and OFEN (Swiss Federal Office of Energy) for past encouragement and support.

Materials for fuel-cell technologies

Brian C. H. Steele* & Angelika Heinzel†‡

*Centre for Ion Conducting Ceramics, Department of Materials, Imperial College, London SW7 2BP, UK (e-mail: b.steele@ic.ac.uk)

†Fraunhofer Institute for Solar Energy Systems, Heidenhofstrasse 2, 79110 Freiburg, Germany

‡Present address: Fachgebiet Energietechnik, Universität Duisburg, Lotharstr. 1-21, 47057 Duisburg, Germany (e-mail: a.heinzel@uni-duisburg.de)

Fuel cells convert chemical energy directly into electrical energy with high efficiency and low emission of pollutants. However, before fuel-cell technology can gain a significant share of the electrical power market, important issues have to be addressed. These issues include optimal choice of fuel, and the development of alternative materials in the fuel-cell stack. Present fuel-cell prototypes often use materials selected more than 25 years ago. Commercialization aspects, including cost and durability, have revealed inadequacies in some of these materials. Here we summarize recent progress in the search and development of innovative alternative materials.

The successful conversion of chemical energy into electrical energy in a primitive fuel cell was first demonstrated¹ over 160 years ago. However, in spite of the attractive system efficiencies and environmental benefits associated with fuel-cell technology, it has proved difficult to develop the early scientific experiments into commercially viable industrial products. These problems have often been associated with the lack of appropriate materials or manufacturing routes that would enable the cost of electricity per kWh to compete with the existing technology, as outlined in a recent survey².

The types of fuel cells under active development are summarized in Fig. 1. The alkaline fuel cell (AFC), polymeric-electrolyte-membrane fuel cell (PEMFC) and phosphoric-acid fuel cell (PAFC) stacks essentially require relatively pure hydrogen to be supplied to the anode. Accordingly, the use of hydrocarbon or alcohol fuels requires an external fuel processor to be incorporated into the system. This item not only increases the complexity and cost of the system, but also decreases the overall efficiency as indicated in Fig. 2. In contrast, molten-carbonate fuel cells (MCFCs) and solid-oxide fuel cells (SOFCs) operating at higher temperatures have the advantage that both CO and H₂ can be electrochemically oxidized at the anode. Moreover, the fuel-processing reaction can be accomplished within the stack, which enables innovative thermal integration/management design features to provide excellent system efficiencies (~50%).

Although the introduction of a 'hydrogen economy' might seem an attractive scenario, its implementation is beset with formidable technical and economic difficulties. The cheapest technology for the large-scale production of hydrogen is the steam reforming of natural gas, which produces significant emissions of greenhouse gases³. The topic of hydrogen storage is addressed in the accompanying review by Schlappbach and Züttel (see pages 353–358); unless there is a breakthrough in the production of hydrogen and the development of new hydrogen-storage materials, the concept of a 'hydrogen economy' will remain an unlikely scenario. In this article, therefore, we assume that fuel cells have to be designed for operation on hydrocarbon or alcohol fuels to ensure that the technology is able to penetrate the relevant major markets. Otherwise fuel-cell technology will be confined to restricted niche activities where hydrogen might be a commercial option, such as city bus fleets. Clearly the choice of fuel is a

further complication in the factors influencing the commercialization of fuel cells.

Constraints on material selection

Materials selection for a commercial product involves an iterative design process that eventually becomes specific to the particular product and application. However, it is possible to make a few general statements about the selection of materials for fuel cells. The combined area-specific resistivity (ASR) of the cell components (electrolyte, anode and cathode) should be below 0.5 $\Omega \text{ cm}^2$ (and ideally approach 0.1 $\Omega \text{ cm}^2$) to ensure high power densities, with targets of 1 kW dm⁻² and 1 kW kg⁻¹ often mentioned for transport applications. High power densities are also important to reduce costs, as the amount of material per kW is thus minimized. These topics, and considerations of cell efficiencies, are summarized in Box 1.

The need to minimize cell resistivities has a major impact on the selection and processing of the cell components. Cost-effective fabrication of porous electrode structures was achieved for the first time only about 40 years ago. The electrolyte, gaseous reactants, electrocatalyst and current collector have to be brought into close contact within a confined spatial region termed the triple-phase-boundary interface. For the low-temperature systems, the introduction of hydrophobic polytetrafluoroethylene (PTFE or Teflon) greatly simplified the fabrication of porous, liquid-resistant gas-diffusion structures. Metal or carbon powders (or porous carbon papers) provided the electronic pathways, and to further reduce the ASR of the electrode a metallic wire mesh or screen was usually incorporated into the structure. Further improvements in performance were obtained during the 1960s by depositing small crystallites (2–5 nm) of the electrocatalyst (usually platinum or Pt alloys) onto carbon powder or paper. In retrospect, this accomplishment was probably the first manifestation of an engineered nanostructure, and it is not surprising that its implementation more than 40 years ago was so difficult.

High ionic conductivities (>1 S cm⁻¹) associated with the liquid KOH, phosphoric acid and molten carbonate electrolytes ensured that, with appropriate design strategies, the ASR values of these components can be small. Although exhibiting lower specific ionic conductivity values, the Nafion membrane used in the PEMFC system can be fabricated relatively easily as a thick film (100 μm) to produce satisfactory ASR values, provided the water content of the film is controlled under the dynamic conditions of cell operation. In contrast, it has been, and continues to be,



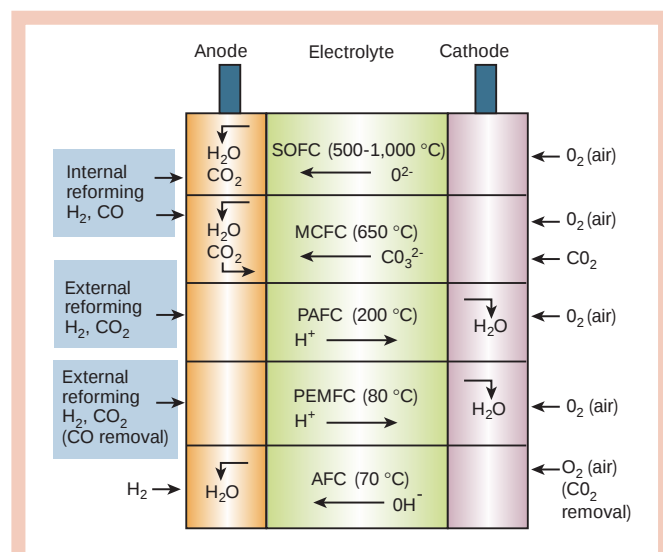


Figure 1 Summary of fuel-cell types. The oxidation reaction takes place at the anode (+) and involves the liberation of electrons (for example, $\text{O}^{2-} + \text{H}_2 = \text{H}_2\text{O} + 2\text{e}^-$ or $\text{H}_2 = 2\text{H}^+ + 2\text{e}^-$). These electrons travel round the external circuit producing electrical energy by means of the external load, and arrive at the cathode (–) to participate in the reduction reaction (for example, $\frac{1}{2}\text{O}_2 + 2\text{e}^- = \text{O}^{2-}$ or $\frac{1}{2}\text{O}_2 + 2\text{H}^+ + 2\text{e}^- = \text{H}_2\text{O}$). It should be noted that as well as producing electrical energy and the reaction products (for example, H_2O and CO_2), the fuel-cell reactions also produce heat. The reaction products are formed at the anode for SOFC, MCFC and AFC types, and at the cathode for PAFC and PEMFC types. This difference has implications for the design of the entire fuel-cell system, including pumps and heat exchangers. To maintain the composition of the electrolyte component in the MCFC system, CO_2 has to be recirculated from the anode exhaust to the cathode input. Additionally, the composition of the polymeric-membrane electrolyte has to be carefully controlled during operation by an appropriate ‘water management’ technology.

difficult to scale-up thick-film technologies to provide cost-effective ceramic solid-oxide electrolyte components with required thicknesses in the range 10–30 μm . Usually the thick-film electrolyte has to be sintered dense at temperatures approaching 1,400 °C; this requires a porous ceramic substrate, which is often the anode or cathode material. The substrate material has to be carefully selected to avoid reaction with the electrolyte, and/or thermal expansion mismatch, during the high-temperature sintering process. The incorporation of a relatively weak, brittle structural component in SOFC stacks is at present restricting the application of SOFC systems to those situations that do not demand rapid temperature fluctuations. In this respect the recent development² of sintering procedures below 1,000 °C, which should allow the use of metal substrates, represents a significant advance that will enable the development of more rugged SOFC systems.

Another important component in a fuel-cell stack is the impermeable electronic conducting bipolar plate. This has the dual function of distributing the fuel and air to the anode and cathode, respectively, as well providing the electrical contact between adjacent cells. The corrosive acidic conditions prevailing in the PEMFC and PAFC systems severely restricts the choice of bipolar plate material and at present graphite is usually selected. However, alternative materials or manufacturing methods are mandatory if these systems are ever to attain the target costs. Major research and development (R&D) programmes are examining the behaviour of alternative carbon-based materials produced by injection moulding, or coated stainless steels. For the high-temperature systems (MCFC and SOFC) operating in the temperature range 500–750 °C, appropriate stainless steel compositions can be specified which seem to satisfy the technical and economic constraints. But for SOFCs operating at higher temperatures (800–1000 °C), alternative, more expensive bipolar plate materials have to be specified, which at present incur significant cost penalties.

Additional constraints influencing material selection arise from reliability and durability issues. For transport applications, minimal values of performance degradation (for example, 0.1% over 1,000 h) are required for projected operational lifetimes of 5,000 h. But for stationary applications — for example, distributed CHP (combined heat and power) systems — a similar degradation rate must extend over a period of at least 40,000 h (5 years). These different lifetime targets seem to be introducing problems for PEMFC prototype CHP systems, as the stack components were developed originally for transport applications.

A fuel-cell system also incorporates relevant balance-of-plant items such as pumps, valves, heat exchangers and piping. Although these important components comprise at least half the cost of a fuel-cell system, we will not consider them further here, except to note that for the PAFC system, many of which have now been operated for periods approaching 30,000 h, the main source of system breakdown has been balance-of-plant items. External fuel processors (reformers) are also the subject of intensive development around the world, and a variety of innovative compact reformers using diffusion-bonded printed-circuit components or micro-channel designs⁴ also illustrate the impact of materials technology on this aspect of fuel-cell systems.

For more than four decades now, reliable, efficient, fuel-cell systems incorporating AFC stacks have proved their worth in the Apollo spacecraft and space shuttles. Excellent electrode kinetics, when operating on pure hydrogen and oxygen, are an attractive feature of this system. But for terrestrial applications, the additional economic restraints, which include the need to replace hydrogen by cheaper hydrocarbon or alcohol fuels, have provided severe problems for materials selection and the associated fuel-processing technology. After 20 years development, the Elenco consortium abandoned attempts in 1996 to develop a bus powered by an AFC system. Although Zevco have purchased the technical rights, it is anticipated that market penetration of AFC systems will be small, in spite of a recent publication⁵ advocating the use of AFC systems with ammonia as the source of hydrogen fuel.

Approximately 200 PAFC co-generation units (International Fuel Cells (IFC) PC25 systems, delivering 200 kW) have been installed around the world, and have exhibited excellent reliability. However, the commercial future of this system is possibly in jeopardy as the manufacturers (IFC and Japanese companies) have been unable to reduce the capital cost sufficiently below US\$3,000 per kW_e as originally forecast^{6,7}. Most observers⁸ believe that for initial market entry the target cost per kW_e must be reduced to around US\$1,000 per kW_e, falling to below US\$500 per kW_e with volume production. Accordingly, we focus here on materials aspects of the PEMFC, MCFC and SOFC systems, which at present still appear to present opportunities to exploit their potential.

It is important to note that the materials currently being used in PEMFC, MCFC and tubular SOFC prototype demonstration units essentially remain the same as those selected at least 25 years ago^{9,10}. Although innovative fabrication and processing routes have improved the attributes (for example, lower cost and lower Pt loadings) of these materials, it is only in the past five years that system engineering and commercialization issues have highlighted the inadequacies of some of the materials originally selected. As indicated in the next two sections, it is these issues that are now driving the development of alternative materials, particularly for the PEMFC and intermediate temperature (IT)-SOFC stacks.

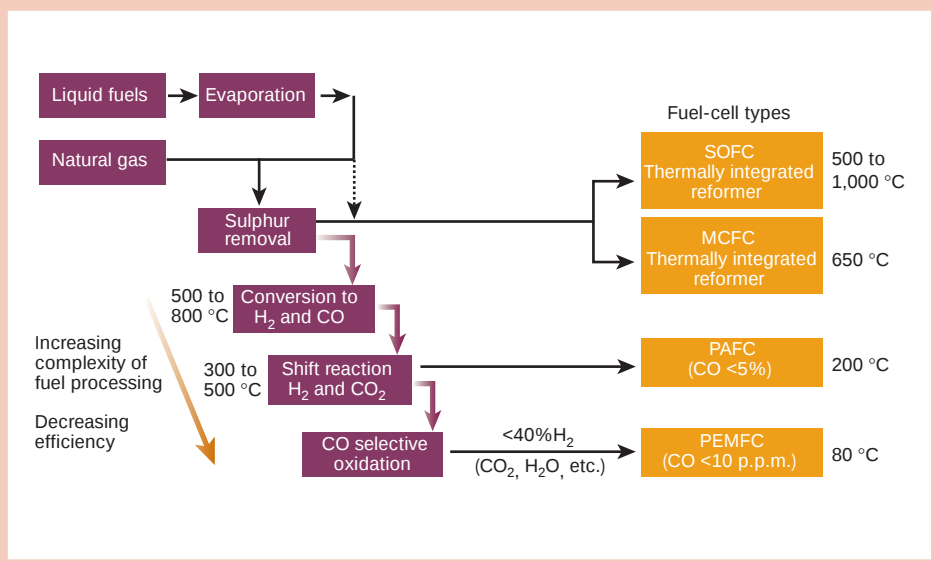
Polymeric-electrolyte-membrane fuel cells

The most important materials under development for PEMFC stacks are construction materials for the cell frames, bipolar plates, electrocatalysts for the fuel and air electrode, and the ion conducting membrane.

Depending on the fuel to be used in the PEM cell, the requirements for these materials are completely different. The simplest case is the operation with pure hydrogen and oxygen or air. Cells with high

Figure 2 Fuel-cell types and fuel processing.

Although selected fuels can be introduced directly into the anode compartment of the high-temperature fuel cells (SOFC and MCFC), better thermal management of the stack can usually be achieved by having separate reformer compartments that are thermally integrated within the stack to produce a mixture of fuel and syngas (H_2 and CO). For the lower-temperature fuel cells (PAFC and PEMFC), external reformers are required. Some of the fuel has to be consumed in these external reformers to maintain the operating temperature. Moreover, dilution of the H_2 fuel reduces performance of the cells, resulting in significant efficiency losses compared with operation on pure H_2 . It should be noted that the AFC stack cannot be operated on reformat fuels because of the presence of CO_2 in these gases.



power density and very low degradation are already state of the art. The main requirement for the future is to achieve reduction in the capital cost of the system by attention to materials selection and fabrication, and also by scaling-up the production volume. This target will probably cause a complete review of all the materials used up to now. The second possibility is to operate the PEM cell with a reformat fuel. In that case, inert gases and CO traces are present in the fuel. This is a challenge for the fuel electrode in particular, and a CO -tolerant catalyst is required. The most difficult option is the direct methanol fuel cell (DMFC). The methanol electrode also needs a CO -tolerant catalyst, as adsorbed CO species are formed during the electro-oxidation of methanol. In addition, the small polar methanol molecule behaves in a way similar to water and readily permeates through existing membrane materials. This behaviour leads to a loss in fuel and also to the formation of a mixed potential at the air electrode. The development of innovative membrane materials could remove this disadvantage. As the tolerance to CO by the electrocatalysts is strongly dependent on temperature, an alternative membrane with a better temperature stability is also another

important R&D target. However, an operation temperature above the boiling point of water requires a completely different type of membrane material, as no liquid water will be present under these circumstances for the hydrated protonic-conduction mechanism.

Bipolar plates

Much effort is being expended on the development of cost-effective materials for the bipolar plates. With respect to corrosion resistance, graphite materials are preferred. However, the conductivity of graphite materials is much less than that of metallic materials. Some conductivity values are: C polymers, $\sim 1 \text{ S cm}^{-1}$; graphite, 10^3 S cm^{-1} ; gold, $45,000 \times 10^3 \text{ S cm}^{-1}$; Fe alloys, $5,300 \times 10^3 \text{ S cm}^{-1}$; Ti, $2,400 \times 10^3 \text{ S cm}^{-1}$. For bipolar plates, polymer/graphite compounds are developed with at least 10 S cm^{-1} , reducing the resistivity of the bipolar plate well below the resistivity of the membrane. In addition, the fabrication costs of graphite plates incorporating gas-distribution channels are high, making such components too expensive. Moreover, as graphite materials are porous, a binder or resin has to be added to produce the necessary impermeability.

Polymeric materials can be machined more easily and cheaply by hot pressing or injection moulding. Polypropylene, for example, can be mixed with graphite to achieve sufficient electrical conductivity. Higher contents of graphite produce better electrical conductivity values, but the associated mechanical properties become more undesirable as the increased brittleness not only reduces the toughness, but also makes manufacturing more difficult and expensive. Typical carbon contents range between 50 and 80 weight%, and several groups^{11–14} are optimizing the machining processes in association with optimization of the material.

Another strategy is to use metallic bipolar plates. The gaseous flow structure can easily be fabricated in thin metal foils by pressing, but only a few metals are sufficiently corrosion-resistant in the acidic environment of the membrane. The most promising materials are stainless steels, as the other candidate metals such as titanium, niobium, tantalum and gold (including gold-plated metals) are too expensive. Stainless steels can provide satisfactory performance for several thousand hours. The steel is protected by a passive layer at the cathode side, but the anode side becomes contaminated by corrosion products^{15–18}. Stacks with metallic bipolar plates have been developed by Nuvera and by Siemens.

Electrocatalyst

The second important problem is associated with the electrocatalyst. For operation with pure hydrogen and air, platinum is the most active material. To reduce the cost, nanoparticles of platinum on a carbon

Box 1

Cell efficiencies

The overall cell efficiency η is given by the equation

$$\eta = \eta_g \eta_v \alpha$$

where η_g is the Gibbs efficiency, η_v is the voltage efficiency, and α is the fraction of fuel used.

$$\eta_g = \Delta G / \Delta H = nFE_0 / \Delta H$$

$$\eta_v = E / E_0 = (E_0 - IR_c) / E_0$$

where E_0 is the open circuit voltage, and ΔH is the heat of the overall cell reaction. Thus

$$\eta = nF(E_0 - IR_c)\alpha / \Delta H$$

R_c is the area specific resistivity of the cell components (electrolyte, anode and cathode). Note that the cell structure is often termed the membrane-electrode assembly (MEA) for PEMFC systems, and positive-electrolyte-negative (PEN) for SOFC systems.

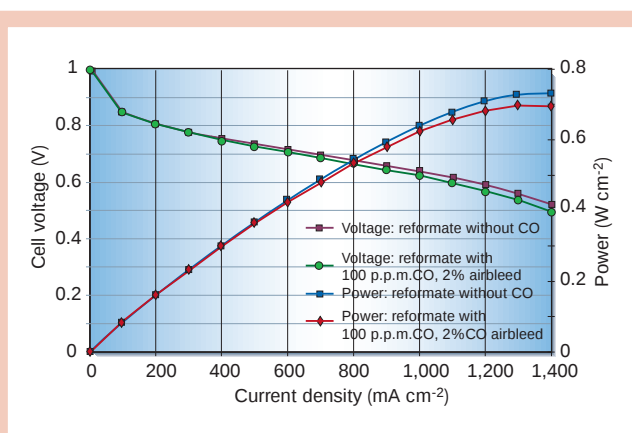


Figure 3 CO tolerance on Pt/Ru anode electrodes. Active cell area, 50 cm²; precious metal loadings, <0.4 mg cm⁻²; temperature, 80 °C; pressure, 3 bar; anode fuel reformat, 60% H₂, 25% CO₂, 15% N₂; cathode, air. (From ref. 56.)

support have been developed, and the reduction in noble-metal content without degrading the cell performance has been, and continues to be, an important R&D activity¹⁹. The platinum loading has been significantly reduced from 2 mg per square centimetre of electrode to values below 0.5 mg cm⁻² without significant impact on performance and lifetime. In the laboratory, even lower platinum loadings have been examined^{20,21}.

For fuels containing traces of CO, or methanol in the DMFC, a CO-tolerant catalyst is required. This remains one of the most challenging tasks for the successful development of commercial PEMFC systems²². For reformat electrodes as well as for methanol oxidation, the removal of adsorbed CO species is the rate-determining step. The oxidation of adsorbed CO on Pt is slow, and is facilitated by adjacent adsorbed OH species. This is the reason why Ru, with its low potential for OH-ads formation, is the most efficient component of the binary catalysts. Pt/Ru and other binary and ternary alloys with these noble metals have been investigated intensively²³, and performance values have increased significantly. The loss in performance is usually expressed in mV for a certain CO content of the fuel, and recent publications show promising results, as indicated, for example, in Fig. 3.

Membranes

Although the US General Electric Company (GE) initiated the development of PEMFCs in the 1950s, it was the introduction of Nafion by DuPont that ensured continuing interest in these systems. Initially Nafion was manufactured²⁴ for membrane cells used in the production of chlorine (chlor-alkali cells). By 1990, Ballard had overcome many of the engineering problems associated with PEMFC systems, and this had stimulated many groups in the United States and Japan to improve the properties of the original Nafion material²⁵. For example, higher ionic conductivities could be attained by selecting perfluorosulphonic acid copolymers with a short pendant group, and it was realized that gas and small-molecule permeability were other important characteristics that had to be improved.

The following properties of polymeric membranes need to be optimized for use in fuel cells: (1) high proton conduction, assured by acid ionic groups (usually SO₃H), depending on sulphonation degree and on thickness of the membrane; (2) good mechanical, chemical and thermal strength requiring the selection of a suitable polymer backbone; mechanical strength for thin membranes can frequently be improved by reinforcement; (3) low gas permeability, which is dependent on material and thickness of the membrane; (4) for DMFC applications low electro-osmotic drag coefficient to reduce methanol crossover.

There is significant interaction between the desired properties of the membrane — high conductivity, low swelling, low gas and

methanol permeability, and stability — and the type of backbone polymer, the degree of sulphonation and the nano-phase separation into hydrophilic and hydrophobic domains (for example, high degrees of sulphonation usually lead to highly conductive membranes, but also to extreme swelling properties). To satisfy these requirements, different approaches have been examined: sulphonated perfluorinated materials with²⁶ and without²⁷ microporous support; sulphonated polyhydrocarbons; acid–base complexes and blends with a surplus of acid ionic groups; and inorganic–organic composite materials with improved thermal stability and better water-retaining properties.

Because of their PTFE-like backbone and relatively low equivalent weight, Nafion and related materials are a favoured option and are commonly used in fuel-cell stacks, but the costs remain high. Therefore, much effort is being applied to the development of cheaper, usually fluorine-free, membrane materials. But hydrocarbons often suffer from an insufficient thermal stability, and so more and more aromatic groups have been introduced into the polymer backbone. Polyarylenes^{28,29} seem to be the most stable molecules among the hydrocarbons. For example, poly(arylene ether sulphone)-based membranes were prepared by sulphonation of commercially available polymers such as Udel and Victrex³⁰. For high ionic conductivity, high sulphonation fractions are desirable, but high sulphonation can lead to extreme swelling even at room temperature. Thus, crosslinking of the polymer chains at the sulphonic acid groups can be included in the synthesis steps to overcome the problems of swelling. However, the long-term stability of these sulphonamide crosslinking bridges remains unproven. Alternative crosslinking methods, such as covalent crosslinking and ionic crosslinking by the introduction of polymeric bases (acid–base blend membranes), have also been examined³¹.

For operation at elevated temperatures, which is desirable for high power density DMFC systems and for reformat fuels with CO levels above 100 p.p.m., the conduction mechanism becomes the dominant issue. In the types of membranes described earlier (hydrated polymers), the proton-conduction mechanism is based on the migration of hydrated protons. Above 100 °C, pressurized operation is required to ensure the presence of liquid water. Phosphoric acid and polymers with immobilized heterocycles exhibit a conduction mechanism relying on structure diffusion³², and can be used at temperatures above the boiling point of water. As phosphoric acid in its liquid form in a porous matrix is well known from PAFC development, new ways to achieve a composite with better properties have been investigated. One polymer material for high-temperature application is polybenzimidazole (PBI), which forms adducts with inorganic acids. Initial publications from Savinell³³ described the use of PBI with phosphoric acid as membrane in fuel-cell experiments. Later developments led to acid–base blend polymers with PBI and sulphonated polyetheretherketones. Reduced methanol permeability and performance data comparable to that of Nafion 112 have been reported^{34,35}. Axiva (part of the Hoechst Celanese group) announced the manufacture of PBI-based membrane materials, but only in exclusive cooperation with partners (Plug Power and Honda).

It should be emphasized that if an alternative membrane material does emerge, considerable R&D will still be necessary to optimize and manufacture the new membrane–electrode assembly (MEA). This development has taken many years for Nafion-type MEAs, although some of the expertise gained may be able to be transferred to the new system.

Fuel cells operating at elevated temperatures

Solid-oxide fuel cells

In contrast to other fuel-cell types a SOFC stack can, in principle, be designed to operate within a wide temperature range (500–1,000 °C). It is necessary, therefore, to select the desired temperature of operation. This, in turn, is influenced by the specific application, the type of fuel and the properties of available solid electrolytes. For example,

if the SOFC stack is to be integrated with a gas turbine, then system requirements indicate that the temperature of the exhaust gas from the stack should exceed about 850 °C. The steam reformation of fuels such as diesel, petrol and propane to produce H_2 and CO gases for the anode also, at present, requires fuel processors operating in excess of 700 °C. But it is the properties of the solid electrolyte that exert the biggest influence on stack design and materials selection, as indicated earlier.

Figure 4 summarizes how the specific ionic conductivity of a selection of solid electrolytes varies with the reciprocal of temperature. By taking a typical value of $0.15 \Omega \text{ cm}^2$ for the maximum ASR value, it is possible to calculate the maximum thickness allowed for a given electrolyte component. For example, a design configuration that specified a self-supported ($\sim 150 \mu\text{m}$), yttria-stabilized zirconia (YSZ) electrolyte would require a temperature of operation greater than 950 °C. Operation at 950 °C poses major problems for planar (flat) SOFC configurations compared to tubular configurations used, for example, in the Siemens–Westinghouse arrangement. These include stability of the electrode–electrolyte interface, selection of the bipolar-plate material, and the optimal material composition and arrangement for the seals that are necessary in planar SOFC stacks. Economic solutions are still sought today.

Owing to these intrinsic problems with the high-temperature planar configuration, there was early interest in tubular designs that eliminated the high-temperature sealing problem. A variety of tubular configurations was initially examined³⁶, but the arrangement eventually adopted by Siemens–Westinghouse has so far proved the most successful. This arrangement uses a 1.5-m porous tubular cathode (La(Sr)MnO₃, or LSM). After deposition of the La(Sr)CrO₃ interconnect strip by plasma spraying, an electrochemical vapour-deposition (EVD) process is used to fabricate an impermeable, thick-film (30–40 μm) electrolyte (YSZ) layer. The cell structure is completed by using a slurry-spray process to deposit the porous Ni-YSZ anode. Although successful from a technical perspective, the EVD process is relatively expensive and efforts are underway to replace this process by an alternative, cheaper fabrication route. However, conventional ceramic routes involving the deposition of YSZ powders and subsequent sintering are constrained by the need to restrict the sintering temperature to below about 1,250 °C to ensure

insignificant reaction between the LSM cathode and electrolyte. Selection of the cathode (LSM) and anode (Ni-YSZ) compositions was established during the 1970s by Westinghouse and ABB, after examining a variety of oxide compositions for long-term compatibility with YSZ at elevated temperatures. This empirical conclusion was later explained in terms of available thermochemical data³⁷.

The Siemens–Westinghouse tubular design remains the most developed SOFC system, and has been evaluated in units generating 25 kW, 100 kW and 200 kW. More recently 200-kW units have been combined with microturbines to provide a system capable of generating electricity around 60% efficiency. Large, multi-megawatt integrated systems are predicted to produce electricity at efficiencies approaching 70%. However, these advances are more concerned with issues of system design rather than the selection and development of new materials.

Intermediate temperature solid-oxide fuel cells

The strategic programmes of large multinational companies (such as Westinghouse, GE and ABB) favoured the development of multi-megawatt, high-temperature SOFC stacks, and these priorities had a strong influence on the development of SOFC designs and materials for two decades from 1970. By 1990, however, it was beginning to be recognized that for smaller SOFC stacks not destined to be integrated with gas turbines, the operating temperature should be lowered as far as possible without compromising the electrode kinetics and internal resistance of the cell. The development of these smaller IT-SOFC stacks for distributed (embedded) CHP units, to produce stand-by power, is also being stimulated by liberalization (deregulation) of electrical supply policies. In addition, many automotive manufacturers are examining whether small SOFC stacks (3–5 kW) can be developed to supply the electrical power for auxiliary functions such as air conditioning in vehicles.

Examples of the most appropriate solid-electrolyte composition for operation at intermediate temperatures (500–750 °C) can be identified from Fig. 4. If we once again assume that the electrolyte component should not contribute more than $0.15 \Omega \text{ cm}^2$ to the total cell ASR, then for a thick-film thickness (L) of 15 μm , the associated specific ionic conductivity (σ) of the electrolyte should exceed $10^{-2} \text{ S cm}^{-1}$ ($\sigma = L/\text{ASR} = 0.0015/0.15$). Examination of Fig. 4 indicates that the ionic conductivity of YSZ attains this target value around 700 °C, and for $\text{Ce}_{0.9}\text{Gd}_{0.1}\text{O}_{1.95}$ (CGO) the relevant temperature is 500 °C. The use of thinner electrolyte films would allow the operating temperature to be lowered. But at present it seems that the minimum thickness for dense impermeable films that can be reliably mass produced using relatively cheap ceramic fabrication routes is around 10–15 μm . The use of a thick-film electrolyte requires this component to be supported on an appropriate substrate. As the substrate is the principal structural component in these cells, it is necessary to optimize the conflicting requirements of mechanical strength and gaseous permeability.

An IT-SOFC configuration that seeks to retain the specific advantages of both the tubular and planar arrangements is being developed by Rolls-Royce³⁸. This integrated planar-stack concept incorporates multi-cell assemblies connected in series and supported by a ceramic substrate, and has many similar features to the original Westinghouse tubular design³⁹.

Most development work on planar IT-SOFC systems has involved thick-film YSZ electrolytes, and so far most groups have used anode (Ni-YSZ) substrates, which allow the electrolyte powder to be sintered to a dense film around 1,400 °C. One of the problems associated with using porous, composite Ni-YSZ substrates is their relatively poor thermal-expansion compatibility with the YSZ thick film. Accordingly, several groups are examining porous substrates based on Ni-Al₂O₃ or Ni-TiO₂ compositions, with thin interfacial anodic regions incorporating Ni, YSZ and/or doped CeO₂. Although replacement of the YSZ can provide better thermal-expansion compatibility, problems still remain over the volume changes associated

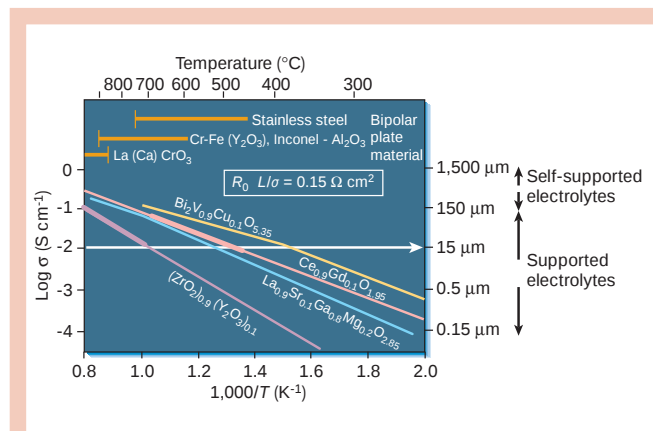


Figure 4 Specific conductivity versus reciprocal temperature for selected solid-oxide electrolytes. To ensure that the total internal resistance (that is, electrolyte + electrodes) of a fuel cell is sufficiently small, the target value for the area specific resistivity (ASR) of the electrolyte is set at $0.15 \Omega \text{ cm}^2$. Films of oxide electrolytes can be reliably produced using cheap, conventional ceramic fabrication routes at thicknesses down to $\sim 15 \mu\text{m}$. It follows that the specific conductivity of the electrolyte must exceed $10^{-2} \text{ S cm}^{-1}$. This is achieved at 500 °C for the electrolyte $\text{Ce}_{0.9}\text{Gd}_{0.1}\text{O}_{1.95}$, and at 700 °C for the electrolyte $(\text{ZrO}_2)_{0.9}(\text{Y}_2\text{O}_3)_{0.1}$. Although the electrolyte $\text{Bi}_2\text{V}_{0.9}\text{Cu}_{0.1}\text{O}_{5.35}$ exhibits higher conductivities, it is not stable in the reducing environment imposed by the fuel in the anode compartment of a fuel cell.

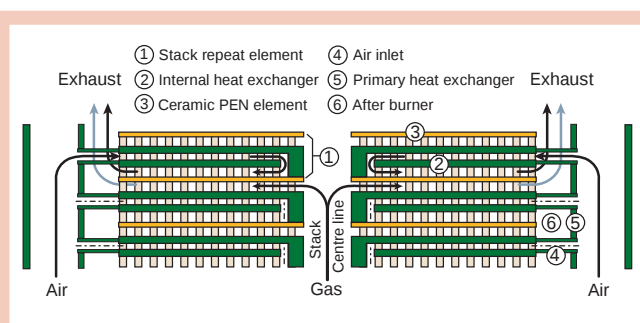


Figure 5 Schematic view of the Sulzer Hexis micro-CHP stack for residential applications. Small (1–5 kW) micro-CHP units are being beta tested for residential applications. The combination of small-scale joint production of electricity and heat is more efficient than the combination of conventional central power stations and decentralized heating equipment. Moreover, security of supply for domestic electrical equipment (for example, central heating pumps and computers) is a major advantage, and part of the emerging trend for distributed power.⁵⁷

with the reduction and oxidation of the Ni component. As the porous substrate–electrolyte films are usually co-fired in air around 1,400 °C, nickel is present as NiO, which has to be carefully reduced by the fuel during the initial heating cycle of the assembled stack. Additionally, operating procedures have to be specified to prevent the nickel re-oxidizing when SOFC stacks are cooled down in the absence of fuel flowing through the anode compartment (of note here is the use of forming gas, N_2/H_2 , to protect the Ni-YSZ anode in the Siemens–Westinghouse tubular configuration).

Most IT-SOFC developers are using metallic bipolar plates. Often a ferritic stainless steel is specified because of the low ($12.5 \times 10^{-6} K^{-1}$) thermal-expansion coefficients of these alloys. Moreover, by using compositions stabilized with Nb–Ti, excellent electronic interfacial contacts can be maintained⁴⁰ between the cell components for extended periods. Providing appropriate precautions are followed, many R&D laboratories have reported good performance values for IT-SOFC stacks incorporating the following positive–electrolyte–negative (PEN) components: anode-supported thick-film YSZ electrolytes, LSM-YSZ cathodes, and stainless steel bipolar plates.

To minimize sealing requirements, many IT-SOFC stacks have adopted a circular design in which the fuel and air are introduced by means of an appropriate manifold at the centre of the PEN structure. Arrangements are made to distribute the air and fuel gases over the cathode and anode, and the flow rates are adjusted to ensure almost complete conversion of the fuel by the time it reaches the stack periphery. Unreacted fuel and air are then combusted without large temperature changes. These design features minimize sealing problems and allow limited thermal cycling. Examples are provided by the Sulzer Hexis micro-CHP configuration designed for residential accommodation (Fig. 5), and the auxiliary power unit (APU) prototype (Fig. 6) constructed by the Delphi–BMW–Global Thermoelectric consortium for incorporation into vehicles. But at present the heating and cooling rates cannot exceed ~500 °C per hour, owing to the development of stresses associated with thermal-expansion mismatch, and to the brittle glass and ceramic seals. Although this restriction may not be too severe for larger CHP systems (>100 kW), it is not satisfactory for smaller systems (1–10 kW) designed for micro-CHP and APU applications. Further R&D is still required to produce more rugged IT-SOFC stacks. Commercial units can be expected in the next five years once reliability and cost requirements (<US\$1,000 per kW) have been effectively demonstrated.

IT-SOFC stacks incorporating alternative components

Although YSZ is still the favoured electrolyte material for SOFC stacks, selection of this material is not without problems and research

continues into the long-term evaluation of scandia-doped ZrO_2 , which can exhibit higher ionic conductivities than the traditional YSZ material.

In principle, the use of ceria-based electrolytes such as CGO should allow the cell operating temperature to be lowered to around 500 °C (see Fig. 4). But perceived problems associated with PEN structures incorporating ceria-based electrolytes have restricted investment in this technology. It is well known that, at elevated temperatures, Ce^{4+} ions can be reduced to Ce^{3+} under the fuel-rich conditions prevailing in the anode compartment. The associated electronic conductivity (and deleterious lattice expansion) produces an internal short circuit in the PEN structure, which can significantly degrade the efficiency and performance of cells. However, if the operating temperature is lowered to around 500 °C, then the electronic conductivity is small, and can be neglected under typical operating conditions of the cell^{41,42}.

Another significant difficulty that has restricted exploitation of the attractive properties of CGO at 500 °C has been the need to develop alternative cathode compositions that function effectively at lower temperatures. Recent developments in this area have been surveyed by Ralph *et al.*⁴³, and there are indications⁴⁴ that appropriate materials for composite cathodes can be fabricated which exhibit small over-potentials at 500 °C (for example, 0.15 V at 1 A cm^{-2}). Composite anodes such as Ni-CGO also provide adequate performance at 500 °C for simulated syngas fuels, indicating that IT-SOFC stacks at 500 °C are now a viable option.

Operation at intermediate temperatures is stimulating R&D effort into alternative anode compositions to replace the established Ni-YSZ anode. One strategy⁴⁵ is to develop electronic conducting oxides that are redox stable, to avoid the use of Ni. This metal can catalyse carbon deposition under certain operating conditions, and also forms NiO, which is accompanied by a deleterious volume expansion, when the anode compartment becomes too oxidizing. Another approach is concerned with the identification of alternative anodes that allow the direct anodic oxidation of hydrocarbons. The claims made in recent publications^{46,47} remain controversial⁴⁸, but there is little doubt that this topic will remain a fruitful area for further investigations.

Operation at 500 °C allows the use of compliant high-temperature gaskets in place of rigid, brittle glass or ceramic seals, thus permitting greater design flexibility for the stack configuration. At Imperial College, London, researchers have taken advantage of the fact that the thermal-expansion coefficient of CGO and ferritic stainless steel are virtually identical ($12.5 \times 10^{-6} K^{-1}$), so that the thick-film PEN structure can be supported on a porous stainless steel foil. These metal-supported PEN structures are robust, and should withstand the rapid temperature cycles expected during operation of small IT-SOFC stacks.

Another electrolyte, doped $LaGaO_3$ (LSGM), is also attracting much attention for IT-SOFC applications. Although its conductivity is slightly smaller (see Fig. 4) than CGO at 500 °C, its ionic domain is wider and it could be more appropriate to use this electrolyte at temperatures around 600 °C, where the reduction of Ce^{4+} in CGO becomes significant. It has been difficult to fabricate pure single-phase ceramic electrolytes, and second phases such as $SrLaGa_3O_7$ and $La_4Ga_2O_9$ are often detected in the grain boundaries. Whether these phases are responsible for the enhanced reactivity of LSGM, or whether it is an intrinsic property of LSGM, are questions that require urgent. Moreover, the preferred composition, $La_{0.9}Sr_{0.1}Ga_{0.8}Mg_{0.2}O_3$, does not seem to be stable at lower temperatures (P. Majewski, personal communication). Although research continues into the synthesis of alternative oxygen-ion conducting electrolytes, it has proved difficult to prepare alternative materials with an appropriate combination of properties that can displace the traditional fluorite compositions involving ZrO_2 and CeO_2 .

Experiments involving single-compartment SOFC fuel cells have been reported. In this configuration a mixture of the fuel and air flows

over both electrodes of a ceramic cell. One electrode incorporates an electrocatalyst optimized for the reduction of oxygen, whereas the second electrode is designed to catalyse the fuel oxidation. This configuration eliminates the need for cell sealing and excellent current–voltage (I – V) characteristics have been reported⁴⁹. But the published data need to be interpreted with caution as it seems that the electrodes also catalyse the external oxidation of the fuel. The heat generated by these surface reactions can significantly raise the temperature of small samples, so that the reported data probably refer to I – V values around 100 °C higher than those stated. Parasitic chemical oxidation of the fuel will of course significantly reduce the overall system efficiency, and technological exploitation of this configuration seems unlikely.

Very high performances have been claimed⁵⁰ at 500 °C using composite electrolytes incorporating CGO and molten salts. But the long-term stability of such mixtures in a fuel-cell environment must be doubtful and these claims require further independent scrutiny.

Investigations also continue into ceramic proton conductors. The composition $\text{BaZr}_{0.9}\text{Y}_{0.1}\text{O}_{2.95}$, for example, can exhibit a protonic conductivity that approaches the oxygen-ion conductivity of CGO at 500 °C (that is, $10^{-2} \text{ S cm}^{-1}$)^{51,52}. But solid-state fuel cells incorporating ceramic proton conductors will not be able to electrochemically oxidize CO, and do not, at present, appear to offer any advantages over the oxygen-ion conducting electrolytes in this temperature region. Interesting results have also been reported⁵³ for the solid acid CsHSO_4 at 160 °C. Ceramic protonic conductors are more likely to be exploited in chemical engineering applications requiring the separation and generation of hydrogen.

Molten-carbonate fuel cells

Although the materials (Table 1) used for MCFC stack components have essentially remained unchanged² over the past 25 years, significant developments in fabrication technology were introduced during the 1980s. Cost-effective tape-casting techniques now allow the immobilized electrolyte matrix to be manufactured up to a size of 1 m². These manufacturing advances were important as the power

Table 1 Materials used in the MCFC stack

Anode	Ni-Cr
Cathode	NiO (Li doped)
Electrolyte	Li/Na/KCO ₃ (60/20/20)
Matrix*	$\gamma\text{-LiAlO}_2 + \text{Al}_2\text{O}_3$
Bipolar plate	Stainless steel 310
Wet seal	Aluminized stainless steel

*For electrolyte immobilization.

density of MCFC systems operating at 650 °C is relatively low ($\sim 150 \text{ mW cm}^{-2}$), requiring large cell areas to be fabricated. The requirement to recirculate CO₂ from the anode exhaust to the cathode to maintain the composition of the carbonate electrolyte also complicates the balance-of-plant equipment. It follows that the MCFC is only likely to be commercialized only at sizes greater than 100 kW, and originally MCFC plants were conceived as large (multi-megawatt scale), centralized coal-fired power stations.

To develop user confidence it was considered that large plants should be demonstrated as soon as possible, and this led to the construction in the United States of the 1.8-MW Santa Clara system by Energy Research Corporation. Although commissioned in April 1996, evaluation of the plant had to be curtailed owing to problems with the material selected for the stack external manifold. It is now believed that a more appropriate strategy is to develop smaller MCFC systems ($\sim 250 \text{ kW}$) for distributed CHP applications using natural gas. An example of this approach is provided by the 300-kW 'Hot Module' unit, developed by MTU Friedrichshafen (a subsidiary of DaimlerChrysler AG), and it is encouraging to note that the first demonstration module has already been operated successfully for more than 1 year (8,000 h) at an electrical efficiency of 47%. Because of the corrosive nature of the molten salt Li/Na/KCO₃ electrolyte, lifetime issues have always been of concern for the MCFC system. However, a recent appraisal⁵⁴ suggests that most components should attain the target value of 40,000 h, except possibly the coating used to protect the anode structure, which remains the focus of further materials development.

Operating procedures for MCFC systems are influenced by limitations associated with two of the components. In cooling down the hot stack in the absence of a fuel supply, it is necessary to protect the Ni-Cr anode from oxidation by the introduction of an inert gas into the anode compartment. More serious, however, is the inability of the immobilized electrolyte matrix to withstand more than 3 to 5 thermal cycles through the melting point of the molten salt electrolyte. Thus, during stand-by situations the temperature of the MCFC stack must be maintained high enough to prevent solidification of the molten salt electrolyte. Clearly both the MCFC and SOFC systems will initially be competing for the same sub-megawatt CHP market sector. It is expected that both will operate at similar electrical efficiencies, and so the cost of the installed plant will determine which technology eventually has the largest market penetration. It is also interesting to note that both fuel-cell types require procedures to prevent redox reactions damaging the anode, and to restrict thermal cycling of the plant during abnormal operating situations.

Conclusions

Probable applications of fuel cells in the next decade together with a selection of critical materials issues are summarized in Table 2. It is recognized that the capital costs (US\$3,000–10,000 per kW) of prototype fuel-cell systems are too high. Although volume production can be expected to reduce these costs, it may be difficult to attain sufficient market share to justify the investment for mass production while competing against established technology. Although significant niche markets exist, such as the PEMFC system for city buses, many observers believe that a more appropriate strategy is to target those sectors of the market (for example, 1–10 kW generation) where the existing technology is inefficient and displays extremely poor

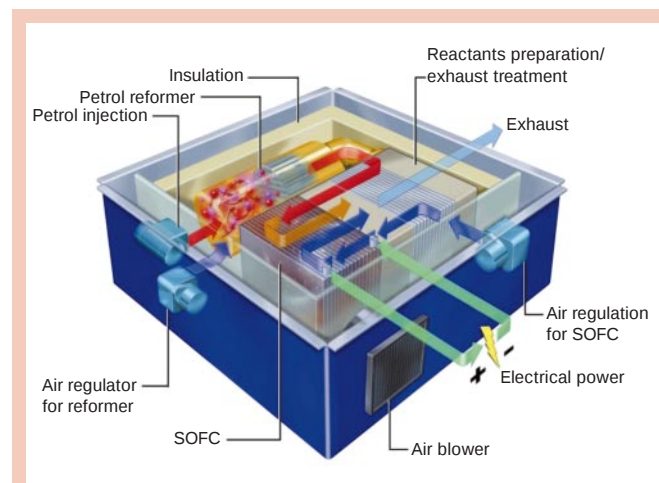


Figure 6 Schematic view of the Delphi–BMW–Global Thermoelectric auxiliary power unit (APU). Depending upon the functions installed in a vehicle, it has been suggested that at least 10% of the output of an internal combustion engine has been used to provide electrical energy. With the introduction of electromagnetic valve trains and smart piezoelectric suspensions this fraction is expected to increase significantly. Furthermore, to maintain air conditioning and refrigeration (in lorries, for example), internal combustion IC engines have to be kept idling even when the vehicle is stationary. Under these low-load conditions the engine is very inefficient and makes significant contributions to environmental pollution. The use of a fuel-cell APU would make a significant contribution to improving air quality and reducing emissions of greenhouse gases.⁵⁸

Table 2 Summary of future R&D requirements for fuel-cell materials

Application	Size (kW)	Fuel cell	Fuel	Critical materials issues
Power systems for portable electronic devices	0.001–0.05	PEMFC DMFC	H ₂ CH ₃ OH	Membranes exhibiting less permeability to CH ₃ OH, H ₂ O
Micro-CHP	1–10	SOFC PEMFC	CH ₃ OH CH ₄ LPG	Novel PEN structures CO-tolerant anodes, novel membranes, bipolar plates
APU, UPS, remote locations, scooters	1–10	SOFC	CH ₄ LPG Petrol	More robust thick-film PENs operating at 500–700 °C
Distributed CHP	50–250	PEMFC MCFC SOFC	CH ₄ CH ₄ CH ₄	CO-tolerant anodes, novel membranes, bipolar plates Better thermal cycling characteristics Cheaper fabrication processes; redox-resistant anodes
City buses	200	PEMFC	H ₂	Cheaper components
Large power units	10 ³ –10 ⁴	SOFC/GT	CH ₄	Cheaper fabrication processes for tubular SOFC system

Abbreviations: UPS, uninterruptible power systems; LPG, liquid petroleum gas; GT, gas turbine. Other abbreviations defined in the text.

part-load performance. This, for example, is the strategy adopted recently by the US Department of Energy for the Solid State Energy Conversion Alliance, which aims to mass produce 5-kW SOFC modular stacks with a target cost of US\$400 per kW. Once this small-scale technology has demonstrated its reliability, and met cost targets, then larger units based on the same technology can be expected to penetrate other sectors of the stationary power and transport market.

Another area receiving increased attention is the development of fuel-cell power sources for a variety of portable electronic devices. As batteries struggle to keep up with the specific power demands of mobile devices, innovative DMFC designs⁵⁵, fabricated using techniques developed for the semiconductor industry, promise energy densities that are between three and five times better than current lithium-ion batteries. The high profit margins and relatively low power demands (~1 W for transmission) associated with cellular telephones, for example, could provide a useful market entry for these small fuel cells, particularly for hybrid systems. □

- Grove, W. R. On voltaic series and the combination of gases by platinum. *Phil. Mag. Ser. 3* **14**, 127–130 (1839).
- Steele, B. C. H. Material science and engineering: the enabling technology for the commercialisation of fuel cell systems. *J. Mater. Sci.* **36**, 1053–1068 (2001).
- Bauen, A. & Hart, J. Assessment of the environmental benefits of transport and stationary fuel cells. *J. Power Sources* **86**, 482–494 (2000).
- Dicks, A. L. & Larminie, J. in *Proc. Fuel Cell 2000* (ed. Blomen, L.) 357–367 (European Fuel Cell Forum, Oberrohrdorf, Switzerland, 2000).
- Kordes, K. et al. Alkaline fuel cells applications. *J. Power Sources* **86**, 162–165 (2000).
- Anahara, R. Total development of fuel cells in Japan. *J. Power Sources* **49**, xi–xiv (1994).
- Whitaker, J. Investment in volume building: the 'virtuous cycle' in PAFC. *J. Power Sources* **71**, 71–74 (1998).
- MacKerron, G. Financial considerations of exploiting fuel cell technology. *J. Power Sources* **86**, 28–33 (2000).
- Kordes, K. & Simader, G. *Fuel Cells and their Applications* (VCH, Weinheim, Germany, 1996).
- Larminie, J. & Dicks, A. *Fuel Cell Systems Explained* (Wiley, Bognor Regis, 2000).
- Borup, R. L. & Vanderborgh, N. E. Design and testing criteria for bipolar plate materials for PEM fuel cell applications. *Mater. Res. Soc. Symp. Proc.* **393**, 151–155 (1995).
- Barbir, F., Joy, G. C. & Weinberg, D. J. in *Proc. Fuel Cell Seminar 2000* 483–486 (Courtesy Associates, Washington DC, 2000).
- Scholta, J., Rohland, B., Trapp, V. & Focken, U. Investigations on novel low-cost graphite composite bipolar plates. *J. Power Sources* **84**, 231–234 (1999).
- Mahlendorf, F., Niemzig, O. & Kreuz, C. in *Proc. Fuel Cell Seminar 2000* 138–140 (Courtesy Associates, Washington DC, 2000).
- Mallant, R., Koene, F., Verhoeve, C. & Ruiter, A. in *1994 Fuel Cell Seminar* 503–506 (Courtesy Associates, Washington DC, 1994).

- Zawodzinski, C., Mahlon, S. & Gottesfeld, S. in *1996 Fuel Cell Seminar* 659–662 (Courtesy Associates, Washington DC, 1996).
- Makkus, R. C., Janssen, A. H. H., de Bruijn, F. A. & Mallant, R. K. A. Use of stainless steel for cost competitive bipolar plates in the SPFC. *J. Power Sources* **86**, 274–282 (2000).
- Davies, D. P., Adcock, P. L., Turpin, M. & Rowen, S. J. Stainless steel as a bi-polar plate material for solid polymer fuel cells. *J. Power Sources* **86**, 237–242 (2000).
- Starz, K. A., Auer, A., Lehmann, Th. & Zuber, R. Characterization of platinum-based electrocatalysts for mobile PEMFC applications. *J. Power Sources* **84**, 167–172 (1999).
- Wilson, M. S., Valerio, J. & Gottesfeld, S. Low platinum loading electrodes for polymer electrolyte fuel cells fabricated using thermoplastic ionomers. *Electrochim. Acta* **3**, 355–363 (1995).
- Uchida, M., Fukuoka, Y., Sugawara, Y., Ohara, H. & Ohta, A. Improved preparation process of very-low-platinum-loading electrodes for polymer electrolyte fuel cells. *J. Electrochem. Soc.* **145**, 3708–3713 (1998).
- Gottesfeld, S. et al. in *Fuel Cell Seminar 2000* 799–802 (Courtesy Associates, Washington DC, 2000).
- McNicol, B. D., Rand, D. A. J. & Williams, K. R. Direct methanol-air fuel cells for road transport. *J. Power Sources* **83**, 15–31 (1999).
- Grot, W., Perfluorinated cation exchange polymers. *Chemie-Ing.-Techn.* **MS260/75** (1975).
- Eisman, G. A. in *Proc. Vol. 86-13* 156–171 (Electrochemical Society, New Jersey, 1986).
- Kolde, J. A., Bahar, B., Wilson, M. S., Zawodzinski, T. A. & Gottesfeld, S. Advanced composite fuel cell membranes. *J. Electrochem. Soc.* **95**, 193–201 (1995).
- Wakizoe, M. & Watanabe, A. in *2000 Fuel Cell Seminar* 27–30 (Courtesy Associates, Washington DC, 2000).
- Huang, R. Y. M. & Kim, J. J. *J. Appl. Polymer Sci.* **89**, 4017, 4029 (1984).
- Zerfaß, T. Thesis, Univ. Freiburg (1998).
- Nolte, R., Ledjef, K., Bauer, M. & Mülhaupt, R. Partially sulphonated poly(arylene ether sulfone)—a versatile proton conducting membrane material for modern energy conversion technologies. *J. Membr. Sci.* **83**, 211–220 (1993).
- Kerres, J. A. Development of ionomer membranes for fuel cells. *J. Membr. Sci.* **185**, 3–27 (2001).
- Kreuer, K. D. On the development of proton conducting polymer membranes for hydrogen and methanol fuel cells. *J. Membr. Sci.* **185**, 29–39 (2001).
- Savinell, R. F. et al. A polymer electrolyte for operation at temperatures upto 200°C. *J. Electrochem. Soc.* **141**, L46–L48 (1994).
- Hasiotis, C. et al. Development and characterization of acid-doped polybenzimidazole/sulfonated polysulfone blend polymer electrolytes for fuel cells. *J. Electrochem. Soc.* **148**, A513–A519 (2001).
- Kerres, J., Ullrich, A., Meier, F. & Haring, T. Synthesis and characterization of novel acid-base polymer blends for application in membrane fuel cells. *Solid State Ionics* **125**, 243–249 (1999).
- Minh, N. Q. & Takahashi, T. *Science and Technology of Ceramic Fuel Cells* (Elsevier, Amsterdam, 1995).
- Yokokawa, H. Phase diagrams and thermodynamic properties of zirconia based ceramics. *Key Eng. Mater.* **154/155**, 37–74 (1998).
- Day, M. J. in *4th European SOFC Forum* (ed. McEvoy, A. J.) 133–140 (European Fuel Cell Forum, Oberrohrdorf, Switzerland, 2000).
- Sverdrup, E. F., Warde, C. J. & Eback, R. L. Design of high temperature solid-electrolyte fuel-cell batteries for maximum power output per unit volume. *Energy Convers.* **13**, 129–141 (1973).
- Dulieu, D. et al. in *3rd European SOFC Forum* (ed. Stevens, P.) 447–458 (European Fuel Cell Forum, Oberrohrdorf, Switzerland, 1998).
- Steele, B. C. H. Appraisal of Ce_{1-x}Gd_xO_{2-y/2} electrolytes for IT-SOFC operation at 500°C. *Solid State Ionics* **129**, 95–110 (2000).
- Xia, C., Chen, F. & Liu, M. Reduced temperature SOFC fabricated by screen printing. *Electrochem. Solid State Lett.* **4**, A52–A54 (2001).
- Ralph, J. M., Schoeler, A. C. & Krumpelt, M. Materials for lower temperature SOFC. *J. Mater. Sci.* **36**, 1161–1172 (2001).
- Doshi, R. et al. Development of SOFCs that operate at 500°C. *J. Electrochem. Soc.* **146**, 1273–1278 (1999).
- Irving, J. T. et al. in *4th European SOFC Forum* (ed. McEvoy, A. J.) 471–477 (European Fuel Cell Forum, Oberrohrdorf, Switzerland, 2000).
- Murray, E. P. & Tsai, T. A direct-methane fuel cell with a ceria based anode. *Nature* **400**, 649–651 (1999).
- Park, S., Craciun, R., Radu, V. & Gorte, R. J. Direct oxidation of hydrocarbons in a SOFC. 1. Methane oxidation. *J. Electrochem. Soc.* **146**, 3603–3606 (1999).
- Primdahl, S. & Mogensen, M. Exchange current densities in mixed conducting SOFC anodes. (Abstr. BS-PO-24, International Society for Solid State Ionics 2001, 8–13 July 2001, Cairns, Australia.) *Solid State Ionics* (in the press).
- Hibino, T. et al. A low operating temperature SOFC in hydrocarbon-air mixtures. *Science* **288**, 2031–2033 (2000).
- Zhu, B. Advantages of intermediate temperature SOFC for tractionary applications. *J. Power Sources* **93**, 82–86 (2001).
- Kreuer, K. D. On the development of proton conducting materials for technological applications. *Solid State Ionics* **97**, 1–15 (1997).
- Bohn, H. G. & Schober, T. Electrical conductivity of the high temperature proton conductor BaZr_{0.9}Y_{0.1}O_{2.95}. *J. Am. Ceram. Soc.* **83**, 768–772 (2000).
- Haile, S. M., Boysen, D. A., Chisholm, C. R. I. & Merle, R. B. Solid acids as fuel cell electrolytes. *Nature* **410**, 910–913 (2001).
- Huijsmans, J. P. P. et al. An analysis of endurance issues for MCFC. *J. Power Sources* **86**, 117–121 (2000).
- Hockaday, R. et al. in *Proc. Fuel Cell 2000* (ed. Blomen, L.) 37–44 (European Fuel Cell Forum, Oberrohrdorf, Switzerland, 2000).
- Starz, K. A., Ruth, K., Vogt, M. & Zuber, R. in *Proc. Int. Symp. Fuel Cells for Vehicles 20-22 November 2000*, 210–215 (Nagoya, Japan, 2000).
- Diethelm, R., Batawi, E. & Honegger, K. in *Proc. Fuel Cell 2000* (ed. Blomen, L.) 203–211 (European Fuel Cell Forum, Oberrohrdorf, Switzerland, 2000).
- Zizelman, J., Botti, J., Tachtler, J. & Wolfgang, S. SOFC auxiliary power unit: a paradigm shift in electric supply for transportation. *Automotive Eng. Int.* **108** (Delphi Suppl.), 14–20 (2000).

Hydrogen-storage materials for mobile applications

Louis Schlappbach*† & Andreas Züttel†

*EMPA, Swiss Federal Laboratories for Materials Research and Testing, CH-8600 Dübendorf, Switzerland

†University of Fribourg, Physics Department, CH-1700 Fribourg, Switzerland (e-mail: andreas.zuettel@unifr.ch)

Mobility — the transport of people and goods — is a socioeconomic reality that will surely increase in the coming years. It should be safe, economic and reasonably clean. Little energy needs to be expended to overcome potential energy changes, but a great deal is lost through friction (for cars about 10 kWh per 100 km) and low-efficiency energy conversion. Vehicles can be run either by connecting them to a continuous supply of energy or by storing energy on board. Hydrogen would be ideal as a synthetic fuel because it is lightweight, highly abundant and its oxidation product (water) is environmentally benign, but storage remains a problem. Here we present recent developments in the search for innovative materials with high hydrogen-storage capacity.

Energy can be stored in different forms: as mechanical energy (for example, potential energy or rotation energy of a flywheel); in an electric or magnetic field (capacitors and coils, respectively); as chemical energy of reactants and fuels (batteries, petrol or hydrogen); or as nuclear fuel (uranium or deuterium). Chemical and electric energy can be transmitted easily because they both involve electronic Coulomb interaction. Chemical energy is based on the energy of unpaired outer electrons (valence electrons) eager to be stabilized by electrons from other atoms. The hydrogen atom is most attractive because its electron (for charge neutrality) is accompanied by only one proton. Hydrogen thus has the best ratio of valence electrons to protons (and neutrons) of all the periodic table, and the energy gain per electron is very high.

Hydrogen is the most abundant element on Earth, but less than 1% is present as molecular hydrogen gas H_2 . The overwhelming majority is chemically bound as H_2O in water and some is bound to liquid or gaseous hydrocarbons. The clean way to produce hydrogen from water is to use sunlight in combination with photovoltaic cells and water electrolysis (see review in this issue by Grätzel, pages 338–344). Other forms of primary energy and other water-splitting processes are also used: the hydrogen consumed today as a chemical raw material (about 5×10^{10} kg per year worldwide) is to a large extent produced using fossil fuels and the reaction of hydrocarbon chains ($-CH_2-$) with H_2O at high temperatures, which produces H_2 and CO_2 . Direct thermal dissociation of H_2O requires temperatures higher than 2,000 °C (>900 °C with a Pt/Ru catalyst).

The chemical energy per mass of hydrogen (142 MJ kg⁻¹) is at least three times larger than that of other chemical fuels (for example, the equivalent value for liquid hydrocarbons is 47 MJ kg⁻¹). Once produced, hydrogen is a clean synthetic fuel: when burnt with oxygen, the only exhaust gas is water vapour, but when burnt with air, lean mixtures have to be used to avoid the formation of nitrogen oxides. Whether hydrogen can be considered a clean form of energy on a global scale depends on the primary energy that is used to split water¹.

The availability of free energy is often unsafe. The mechanical energy of a 1,000-kg car running out of control

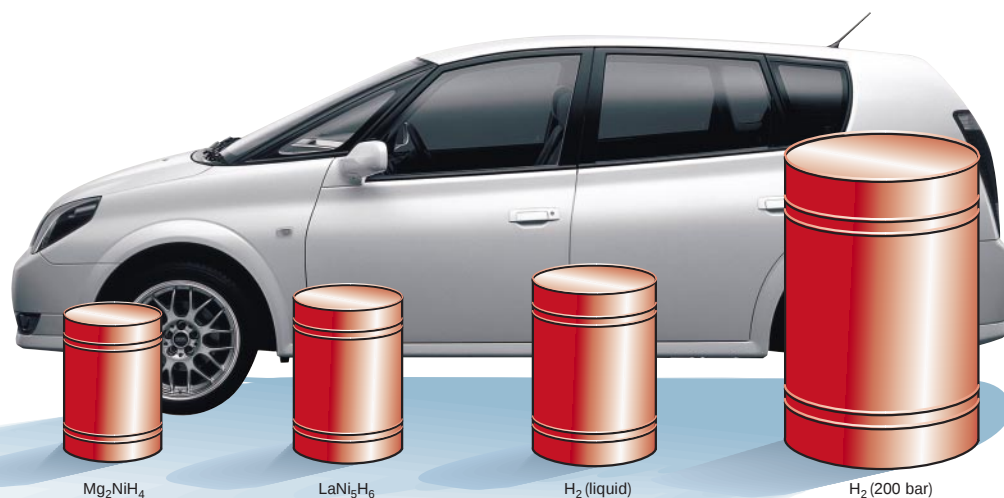
at 40 km h⁻¹ can kill pedestrians. The process of burning hydrogen can be done in an efficient and controlled way to liberate energy at a desirable rate, or in an uncontrolled way with the potential to cause damage. For historical reasons hydrogen has a bad reputation, which is not altogether justified: a more recent analysis² of the *Hindenburg* catastrophe shows that the air ship caught fire because of a highly flammable skin material and not because of the hydrogen gas it contained. The safety of hydrogen relies on its high volatility and non-toxicity.

Today, many scientists and engineers, some companies, governmental and non-governmental agencies and even finance institutions are convinced that hydrogen's physical and chemical advantages will make it an important synthetic fuel in the future. After the successful use of hydrogen for space technology, national hydrogen associations were created and joint ventures started. Examples are fuel-cell development projects, the Shell-GfE-Hydro-Québec joint venture³, the International Hydrogen Energy Association and its technical-economic conferences, and solar hydrogen R&D programmes. But are these ventures aimed at mobility?

There are essentially two ways to run a road vehicle on hydrogen. In the first, hydrogen in an internal combustion engine is burnt rapidly with oxygen from air. The efficiency of the transformation from chemical to mechanical through thermal energy is limited by the Carnot efficiency and is slightly higher for hydrogen–air mixtures (around 25%) than for petrol–air mixtures. When a lean mixture is used, the exhaust gas contains nothing but water vapour; richer mixtures also produce NO_x . In the second method, hydrogen is 'burnt' electrochemically with oxygen from air in a fuel cell, which produces electricity (and heat) and drives an electric engine (see review in this issue by Steele and Heinzel, pages 345–352). Here, the efficiency of the direct process of electron transfer from oxygen to hydrogen is not limited by the Carnot efficiency; it can reach 50–60%, twice as much as the thermal process.

For on-board energy storage, vehicles need compact, light, safe and affordable containment. A modern, commercially available car optimized for mobility and not prestige with a range of 400 km burns about 24 kg of petrol in a combustion engine; to cover the same range, 8 kg hydrogen

Figure 1 Volume of 4 kg of hydrogen compacted in different ways, with size relative to the size of a car. (Image of car courtesy of Toyota press information, 33rd Tokyo Motor Show, 1999.)



are needed for the combustion engine version or 4 kg hydrogen for an electric car with a fuel cell.

Hydrogen is a molecular gas. At room temperature and atmospheric pressure, 4 kg of hydrogen occupies a volume of 45 m³. This corresponds to a balloon of 5 m diameter — hardly a practical solution for a vehicle (Table 1). In the following we will focus on the question of compacting hydrogen, looking at the materials, technology and safety aspects (Fig. 1).

Conventional hydrogen storage

Classical high-pressure tanks made of fairly cheap steel are tested up to 300 bar and regularly filled up to 200 bar in most countries. To store our 4 kg hydrogen still requires an internal volume of 225 litres (about 60 gallons) or 5 tanks of 45 litres each. Novel high-pressure tanks made of carbon-fibre-reinforced composite materials are being developed; these are tested up to 600 bar and filled up to 450 bar for regular use. But they need a special inert inner coating to prevent the high-pressure hydrogen reacting with the polymer. Consequently, another approach is to use hydrogen-inert aluminium tanks and to strengthen them with external carbon-fibre coatings. Spherical containers slightly smaller than 60 cm in diameter would be able to carry our 4 kg, but for practical fabrication a cylindrical shape is preferred.

These high-pressure containers, when full, would contain about 4% hydrogen by mass, but with significant disadvantages: the fuel would be available at a pressure dropping from 450 bar to zero overpressure, so additional pressure control would be essential. High-pressure vessels present a considerable risk — the compression itself is the most dangerous and complicated part. In Japan such vessels are prohibited on the roads in ordinary cars.

Condensation into liquid or even solid hydrogen is, of course, particularly attractive from the point of view of increasing the mass

per container volume. The density of liquid hydrogen is 70.8 kg m⁻³ (70.6 kg m⁻³ for solid hydrogen). But the condensation temperature of hydrogen at 1 bar is -252 °C and the vaporization enthalpy at the boiling point amounts to 452 kJ kg⁻¹. As the critical temperature of hydrogen is -241 °C (above this temperature hydrogen is gaseous), liquid hydrogen containers are open systems to prevent strong overpressure. Therefore, heat transfer through the container leads directly to the loss of hydrogen. Larger containers have a smaller surface to volume ratio than small containers, so the loss of hydrogen is smaller. The continuously evaporated hydrogen may be catalytically burnt with air in the overpressure safety system of the container or collected again in a metal hydride. (Solid hydrogen is a molecular insulating solid; under high pressure it transforms into metallic, possibly even superconducting hydrogen⁴ with T_c of 200–300 °C.)

Cryotechniques for cooling and superinsulated low temperature storage units were developed and proven in space technology. Liquid hydrogen is a fuel in the launching process of the Space Shuttle and in *Ariane*. A Lockheed military-type aircraft and a Tupolev supersonic aircraft have been flown with engines fuelled by liquid hydrogen. BMW have built an automated liquid-hydrogen filling station, and developed and tested several cars running with hydrogen in newly designed vessels to reduce losses by evaporation to below 1.5 mass% per day (BMW, personal communication).

Hydrocarbons with a molecular mass of at least 60 g mol⁻¹ are liquid at room temperature with a density close to 1,000 kg m⁻³. The number of hydrogen atoms per carbon atom can vary in hydrocarbons owing to their ability to change from σ -bonds to π -bonds. Hydrocarbons can be burnt completely by oxidation of carbon into CO₂ and of hydrogen into H₂O; some can also be considered as a liquid storage medium for hydrogen if they can be hydrogenated and dehydrogenated; that is, if their ratio of hydrogen to carbon atoms can be adapted reversibly. Cyclohexane (C₆H₁₂), for example, reversibly desorbs six hydrogen atoms (7.1 mass%) and forms benzene (C₆H₆). Stationary hydrogenation and dehydrogenation under steady-state conditions are managed in numerous chemical plants, but the on-board process under variable conditions is another matter. We do not consider hydrogen storage in NH₃ (5.9 mass%) to be realistic, owing to the corrosive nature of ammonia.

Hydrogen adsorption on solids of large surface area

Hydrogen adsorbs at solid surfaces depending on the applied pressure and the temperature. The variation of attractive surface forces as a function of distance from the surface decides whether van der

Table 1 Physical and chemical properties of hydrogen, methane and petrol

Properties	Hydrogen (H ₂)	Methane (CH ₄)	Petrol (—CH ₂ —)
Lower heating value (kWh kg ⁻¹)	33.33	13.9	12.4
Self-ignition temperature (°C)	585	540	228–501
Flame temperature (°C)	2,045	1875	2,200
Ignition limits in air (Vol%)	4–75	5.3–15	1.0–7.6
Minimal ignition energy (mW s)	0.02	0.29	0.24
Flame propagation in air (m s ⁻¹)	2.65	0.4	0.4
Diffusion coefficient in air (cm ² s ⁻¹)	0.61	0.16	0.05
Toxicity	No	No	High

Waals-type weak physisorption of molecular hydrogen occurs, or whether dissociation and chemisorption of atomic hydrogen takes place. Owing to the attractive forces, the most stable position for an adsorbed molecule is with its centre at about 1 molecular radius from the surface, and the attractive field rapidly diminishes at greater distances. Once a monolayer of adsorbate molecules or atoms has formed, the gaseous species interacts with the liquid or solid adsorbate. Therefore, the binding energy of the second layer of adsorbates is similar to the latent heat of sublimation or vaporization of the adsorbate. Consequently, adsorption at a temperature at or above the boiling point of the adsorbate at a given pressure leads to the adsorption of a single monolayer. For storage purposes, the adsorption of hydrogen has been studied on carbon species only. Other light and reasonably cheap materials of high surface area may prove to be attractive as well.

The condensation of a monolayer of hydrogen on a solid leads to a maximum of $1.3 \times 10^{-5} \text{ mol m}^{-2}$ of adsorbed hydrogen. In the case of a graphene sheet with a specific surface area of $1,315 \text{ m}^2 \text{ g}^{-1}$ as adsorbent, the maximum theoretical concentration is 0.4 H atoms per surface carbon atom or 3.3 mass% hydrogen on sheets with hydrogen atoms on one side. On active carbon with the specific surface area of $1,315 \text{ m}^2 \text{ g}^{-1}$, 2 mass% hydrogen is reversibly adsorbed at a temperature of 77 K (ref. 4). On nanostructured graphitic carbon at 77 K (liquid nitrogen temperature), the reversibly adsorbed quantity of hydrogen correlates with the specific surface area of the sample. It amounts to 1.5 mass% per $1,000 \text{ m}^2 \text{ g}^{-1}$ of specific surface area (Fig. 2), in agreement with the calculation shown in Fig. 3. Nanostructured graphite produced by ball milling for 80 h in a 1-MPa hydrogen atmosphere contains up to 0.95 H atoms per carbon atom or 7.4 mass%. Eighty per cent of this hydrogen desorbs at a temperature of over 600 K. Evidently, moderately high hydrogen absorption can be realized with planar graphitic structures at low temperature^{7,8}.

Are curved structures more attractive? In microporous solids with capillaries which have a width not exceeding a few molecular diameters (the diameter of H_2 is 0.41 nm), the potential fields from opposite walls overlap so that the attractive force acting on hydrogen molecules is greater than that on an open flat surface. The effect of nanotube curvature on the adsorption energy for hydrogen has been investigated theoretically by Stan and Cole⁹. The Feynman (semiclassical) effective potential approximation was used to calculate the adsorption potential and the amount of hydrogen adsorbed on a (13,0) zigzag nanotube. The adsorption potential was found to be 9 kJ mol^{-1} for hydrogen molecules inside the nanotubes at 50 K, about 25% higher than on the flat surface of graphite. The increase arises

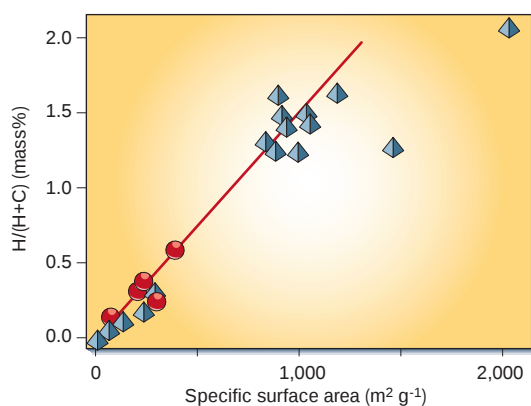


Figure 2 Reversibly stored amount of hydrogen on various carbon materials versus the specific surface area of the samples. Circles represent nanotube samples (best-fit line indicated), triangles represent other nanostructured carbon samples¹⁷.

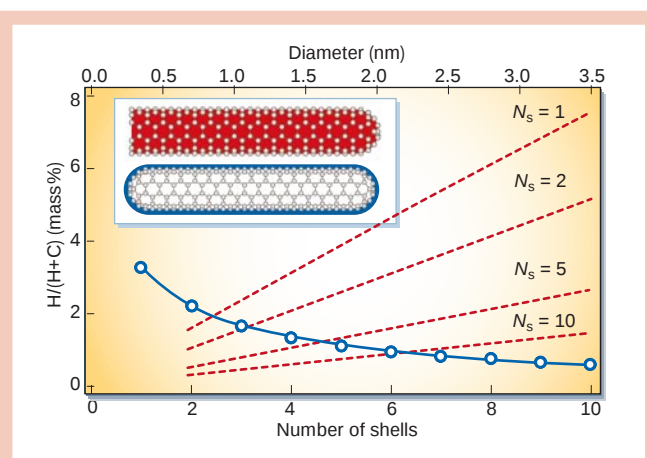


Figure 3 Hydrogen in carbon nanotubes. Calculated storage density in mass% as a function of the number of shells N_s (blue) and as a function of the tube diameter (red). Hydrogen condenses as a monolayer at the surface of the nanotube or condenses in the cavity of the tubes. The condensed hydrogen has the same density as liquid hydrogen at -253°C .

from the curvature of the surface and the related higher number of carbon atoms interacting with the hydrogen molecule. At low temperatures, far more hydrogen can be adsorbed in the tube than on a flat surface; but the ratio decreases strongly with increasing temperature, from 55 at 50 K to 11 at 77 K.

Conflicting results have been published concerning the reversible storage of hydrogen in carbon nanotubes (see refs 8, 9 for detailed reviews). To a large extent the controversy is caused by insufficient characterization of the carbon material used. It is often a mixture of opened and unopened, single-walled and multi-walled tubes of various diameters and helicities together with other carbonaceous species, just a few of which will be analysed in transmission electron micrographs. To our knowledge, the 'fantastic' results of the Northwestern University group¹⁰ (more than 60 mass% hydrogen in specific carbon fibres) have not been reproduced elsewhere. Heben *et al.*^{11,12} still claim to reach 6–8 mass% reversible storage (cooling at 1 bar down to liquid nitrogen temperature) in high-purity single-walled tubes opened by sonication. But Hirscher *et al.*¹³ revealed that a sonication bar made of titanium alloy results in hydrogen-storing titanium alloy particles in the nanotube material; storage was below 1 mass% when stainless steel sonication bars were used. Zuetzel *et al.*¹⁴ showed by thermal desorption mass spectroscopy that after reaction of carbon nanotubes with hydrogen at elevated temperatures, hydrocarbons are desorbed. The large mass increase observed by Chen *et al.*¹⁵ when carbon nanotubes filled with alkaline metals were exposed to hydrogen has been attributed to hydroxide formation¹⁶.

Nijkamp *et al.*¹⁷ have surveyed the storage capacities of a large number of different carbonaceous adsorbents for hydrogen at 77 K and 1 bar. They concluded that microporous adsorbents, for example zeolites and activated carbons, have appreciable sorption capacities. Optimization of sorbent and adsorption conditions is expected to lead to adsorption of $560 \text{ ml STP g}^{-1}$, close to targets set for practical use in vehicles. Zuetzel *et al.*¹⁸ concluded that charging of carbon nanotubes with hydrogen at liquid nitrogen temperature (77 K), or cathodically at ambient conditions, is due to physisorption. The amount of adsorbed hydrogen is proportional to the specific surface area of the carbon nanotube material and limited to 2 mass% for carbon materials (Fig. 2).

Taking all these results together, we consider it to be scientifically interesting and challenging to continue research on the interaction of hydrogen with different and well-characterized carbon nanostructures. Whether a hydrogen-storage material will emerge from it, however, remains an open question.

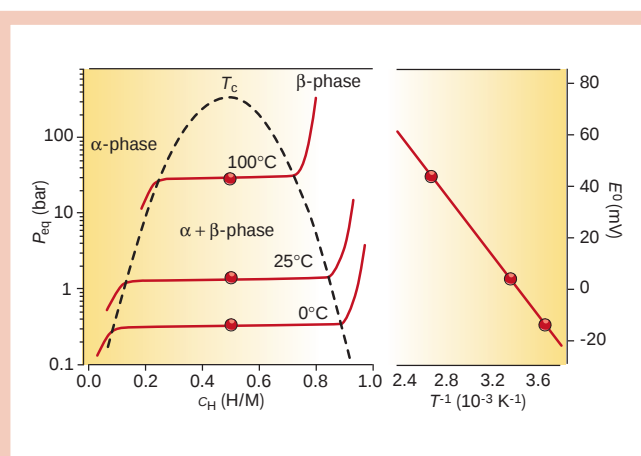


Figure 4 Pressure–concentration–temperature plot and a van't Hoff curve (logarithm of the equilibrium or plateau pressure against the reciprocal temperature); values are for LaNi_5 . The vertical axes indicate the corresponding hydrogen pressure or the equivalent electrochemical potential. From the slope of the van't Hoff plot, experimental values of the enthalpy of hydride formation ΔH can be evaluated. The plateau pressure $p_{\text{eq}}(T)$ as a function of temperature is related to the changes ΔH and ΔS of enthalpy and entropy, respectively, by the van't Hoff equation: $\ln(p_{\text{eq}}/p_{\text{eq}}^0) = (-\Delta H/R)(1/T) + \Delta S/R$.

In an interesting sideline to work¹⁹ — addressing the potential of metallization and superconductivity of hydrogen, rather than its storage — some hydrogen was reported to be sorbed into graphite intercalation compounds made of alkaline metals between graphene sheets. Intercalation of hydrogen alone between the graphene layers in macroscopic structures has never been reported. The dynamic diameter of a free hydrogen molecule ($d = 0.4059$ nm) is larger than the interlayer distance in graphite ($d = 0.3355$ nm), so a significant size mismatch would have to be accommodated upon entry.

Hydrogen storage by metal hydrides

Many metals and alloys are capable of reversibly absorbing large amounts of hydrogen. Charging can be done using molecular hydrogen gas or hydrogen atoms from an electrolyte. Molecular hydrogen is dissociated at the surface before absorption; two H atoms recombine to H_2 in the desorption process. The thermodynamic aspects of hydride formation from gaseous hydrogen are described by pressure–composition isotherms (Fig. 4). The host metal initially dissolves some hydrogen as a solid solution (α -phase). As the hydrogen pressure together with the concentration of H in the metal is increased, interactions between hydrogen atoms become locally important, and we start to see nucleation and growth of the hydride (β) phase. While the two phases coexist, the isotherms show a flat plateau, the length of which determines how much H_2 can be stored reversibly with small pressure variations. In the pure β -phase, the H_2 pressure rises steeply with the concentration. At higher H_2 pressure, further plateaux and further hydride phases may be formed. The two-phase region ends in a critical point T_c , above which the transition from α - to β -phase is continuous. The plateau or equilibrium pressure depends strongly on temperature and is related to the changes ΔH and ΔS of enthalpy and entropy, respectively. As the entropy change corresponds mostly to the change from molecular hydrogen gas to dissolved hydrogen, it is roughly $130 \text{ J K}^{-1} \text{ mol}^{-1}$ for all metal–hydrogen systems under consideration. The enthalpy term characterizes the stability of the metal–hydrogen bond. To reach an equilibrium pressure of 1 bar at 300 K, ΔH should amount to 19.6 kJ mol^{-1} (ref. 20). The operating temperature of a metal hydride system is fixed by the plateau pressure in thermodynamic equilibrium and by the overall reaction kinetics²¹.

Hydrogen is located in the form of atoms, never molecules, on interstitial sites of the host metal lattice. The lattice expands during

hydrogen sorption, often losing some of its high symmetry. As a consequence of the coexistence of the non-expanded α -phase and anisotropically expanded β -phase, lattice defects and internal strain fields are formed, which end in a decrepitation of brittle host metals such as intermetallics. The H atoms vibrate about their equilibrium position, and perform local motions and long-range diffusion.

In terms of electronic structure, the proton acts as an attractive potential to the host metal electrons; electronic bands are lowered in energy and form low-lying bands by hybridization with the hydrogen band, 6–8 eV below the Fermi level. The Fermi level itself is shifted, and various phase transitions (metal–semiconductor, magnetic–non-magnetic, reflecting–transparent, order–disorder) may occur. The metal–hydrogen bond offers the advantage of very high hydrogen density at moderate pressure and desorption of all stored hydrogen at the same pressure.

Which metallic systems are appropriate for hydrogen storage? Many elemental metals form hydrides, for example $\text{PdH}_{0.6}$, rare earth REH_2 , REH_3 or MgH_2 , none of which are in the pressure and temperature range attractive for mobile storage (1–10 bar, 0–100 °C, corresponding to an enthalpy change between 15 and 24 kJ mol^{-1}). The discovery of hydrogen sorption by intermetallic compounds created great hopes and stimulated R&D efforts worldwide. Well-known compounds and their properties are summarized in Table 2 (ref. 22).

Alloys derived from LaNi_5 show some very promising properties, including fast and reversible sorption with small hysteresis, plateau pressure of a few bars at room temperature and good cycling life. The volumetric hydrogen density (crystallographically) of $\text{LaNi}_5\text{H}_{6.5}$ at 2 bar equals that of gaseous molecular hydrogen at 1,800 bar; but advantageously, all the hydrogen desorbs at a pressure of 2 bar. The density for practical purposes is reduced by the packing fraction of LaNi_5 powder, but is still above that of liquid hydrogen (Fig. 5). Storage in these intermetallics allows very safe hydrogen handling. But as lanthanum and nickel are large elements, the proportion of hydrogen in $\text{LaNi}_5\text{H}_{6.5}$ remains below 2 mass%. This is attractive for electrochemical hydrogen storage in rechargeable metal hydride electrodes, reaching a capacity of 330 mA h g^{-1} , and produced and sold in more than a billion AB_5 -type metal hydride batteries per year^{23–25}.

For gaseous hydrogen fuel tanks to be used in vehicles, however, this is not enough. We need some 4–5 mass% (indeed, 6.5 mass% and $62 \text{ kg H}_2 \text{ m}^{-3}$ are the targets of the US Department of Energy). This low mass density is the general weakness of all known metal hydrides working near room temperature (Fig. 6). Of course, many intermetallic compounds and alloys are known that form hydrides with up to 9 mass% hydrogen (such as $\text{Li}_3\text{Be}_2\text{H}_7$; ref. 26) and 4.5 H atoms per metal atom (BaReH_9 ; ref. 27), but they are not reversible within the required range of temperature and pressure.

When reaction kinetics rather than thermodynamic equilibrium conditions limit the hydride formation and decomposition, various physical and chemical pretreatments can be applied. Ball milling has the benefits of reducing the grain size, increasing the defect concentration and shortening the diffusion path. Fluorination treatments²⁸ are well suited to render the surface active, apparently by means of the formation of surface nickel precipitates, the role of which is well known in the hydrogen dissociation process.

Many promising new ideas are under investigation with the clear goal of enhancing the mass density. For example, new families of

Table 2 Intermetallic compounds and their hydrogen-storage properties²²

Type	Metal	Hydride	Structure	mass%	p_{eq}, T
Elemental	Pd	$\text{PdH}_{0.6}$	$Fm\bar{3}m$	0.56	0.020 bar, 298 K
AB_5	LaNi_5	LaNi_5H_6	$P6/mmm$	1.37	2 bar, 298 K
AB_2	ZrV_2	$\text{ZrV}_2\text{H}_{6.5}$	$Fd\bar{3}m$	3.01	10^{-8} bar, 323 K
AB	FeTi	FeTiH_2	$Pm\bar{3}m$	1.89	5 bar, 303 K
A_2B	Mg_2Ni	Mg_2NiH_4	$P6222$	3.59	1 bar, 555 K
Body-centred cubic	TiV_2	TiV_2H_4	b.c.c.	2.6	10 bar, 313 K

alloys are being studied in several Japanese laboratories^{29–31} and are included into the national 'Protium' programme. They are based on vanadium, zirconium and titanium as rather electropositive components combined with 3d and 4d transition metals. Reversible hydrogen-storage capacities approaching 3 mass% around room temperature have been reported.

A higher mass density is reachable only with light elements such as calcium and magnesium. In fact, Mg forms ionic, transparent MgH_2 containing 7.6 mass% hydrogen. But its formation from bulk Mg and gaseous hydrogen is extremely slow, and in thermodynamic equilibrium a plateau pressure of 1 bar requires not room temperature but 300 °C. Different processes have been tried to obtain micro- or nanostructured Mg: precipitations from metal-organic solutions or high-energy ball milling of Mg have proved successful ways to obtain good charging or discharging kinetics at 150 °C, and the thermodynamics evidently are not affected^{26,29–31}. Alloying Mg before the hydride formation is another approach: Mg_2Ni forms a ternary complex hydride Mg_2NiH_4 , which still contains 3.6 mass% hydrogen. The hydride forms fairly rapidly, probably owing to the presence of Ni as catalyst for the dissociation of molecular hydrogen, but thermodynamically it still requires 280 °C for 1 bar hydrogen. The alloys Mg_2Cu , $\text{Mg}_{17}\text{La}_2$ and MgAl , and some other known alloys or intermetallic compounds of Mg, react readily (MgAl after ball milling) with hydrogen and decompose into MgH_2 and another compound or hydride. One example is $\text{Mg}_2\text{Cu} + \text{H}_2$ which decomposes to $\text{MgH}_2 + \text{MgCu}_2$. The reactions are reversible at high temperature.

In addition to the enthalpy change of the hydrogen sorption reaction, we now also have to consider the enthalpy change of the host metal system. Could this make Mg-based storage alloys more attractive? If the transformation of the host metal system that accompanies hydrogen desorption is exothermic itself, the net enthalpy of hydrogen desorption becomes smaller. In the systems studied so far (Mg–Cu, Mg–Al), the temperature for 1 bar equilibrium pressure could be lowered from 300 °C to 280 °C. This is not a large improvement, and it has the penalty that the mass density is thereby reduced as well. Magnesium does not form a binary intermetallic compound with iron, but in the presence of hydrogen it is possible to synthesize the rather stable ternary hydride Mg_2FeH_6 with 5.5 mass% hydrogen²⁶.

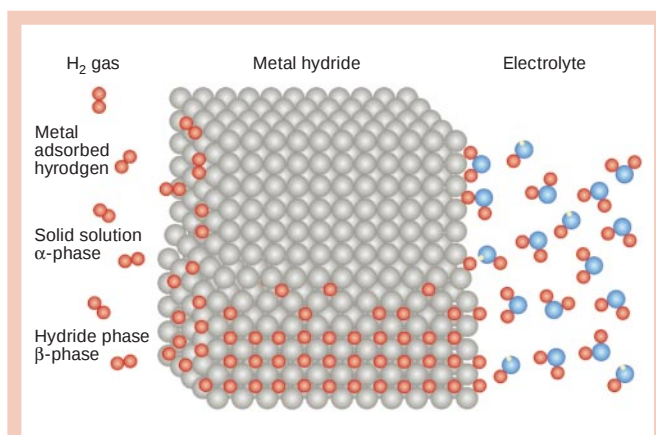


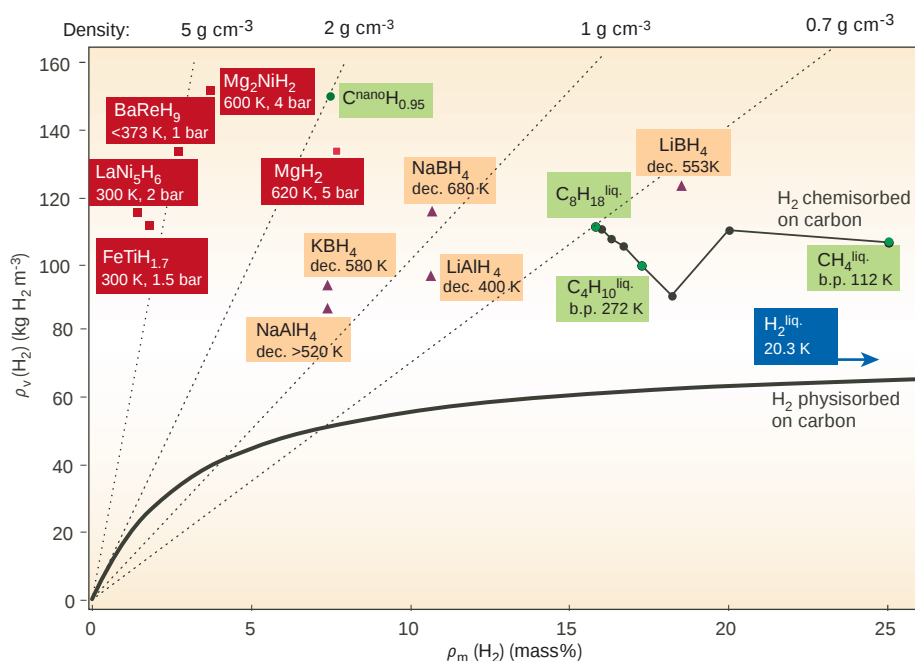
Figure 5 Schematic model of a metal structure with H atoms in the interstices between the metal atoms, and H_2 molecules at the surface. Hydrogen atoms are from physisorbed hydrogen molecules on the left-hand side and from the dissociation of water molecules on the right-hand side.

A different approach is to use composite materials, in which attractive properties of two components are combined to overcome their weaknesses. For instance, magnesium has been ball milled with graphitic carbon or mixed with hydrides showing fast kinetics, such as LaNi_5 or Mg_2Ni . It comes as no surprise that the capacities reached values between those of the components^{5,26,32}.

Alانات and other light hydrides

Some of the lightest elements in the periodic table, for example lithium, boron, sodium and aluminium, form stable and ionic compounds with hydrogen. The hydrogen content reaches values of up to 18 mass% for LiBH_4 . However, such compounds desorb the hydrogen only at temperatures from 80 °C up to 600 °C, and the reversibility of the reaction is not yet clear for all systems. Bogdanovic and Schwickardi³³ showed in 1996 that the decomposition temperature of NaAlH_4 can be lowered by doping the hydride with TiO_2 . The same group showed the reversibility of the reaction

Figure 6 Stored hydrogen per mass and per volume. Comparison of metal hydrides, carbon nanotubes, petrol and other hydrocarbons.



for several desorption/absorption cycles. This is a good example of the potential of such hydrides, which were discovered more than 50 years ago. But several points have to be clarified. First, is the high desorption temperature due to the poor kinetics of the system or due to the thermodynamic stability of the compound? The kinetics can be improved by applying an appropriate catalyst to the system and apparently also by ball milling and the introduction of defects. Second, what are the conditions for a reversible reaction — for example, formation and stabilization of clusters of an intermetallic compound of the remaining metals on desorption? Third, what is the desorption reaction and what are the intermediate reaction products (decomposition in several steps)?

Assuming that these questions can be answered, what would be the role of light hydrides in an on-board fuel cell? The exhaust gas of a fuel cell is water vapour, which could be collected and reused for on-board hydrogen production. The common experiment — shown in many chemistry classes — where a small piece of sodium floating on water produces hydrogen, demonstrates such a process. The sodium is transformed to sodium hydroxide in this reaction. The reaction is not reversible, but the sodium hydroxide could later be removed and reduced in a solar furnace back to sodium. Each sodium atom produces one hydrogen atom, so the corresponding gravimetric hydrogen density of the sodium reaction is slightly more than 4 mass%. Lithium used in the same way would deliver up to 14 mass% of hydrogen. The alkali metals as a hydrogen source are easy to handle, and a car could be refilled within a few minutes. To deliver the necessary 4 kg of hydrogen using the water produced in the fuel cell would take 28 kg of lithium. After using up all the hydrogen the tank would contain 99 kg of lithium hydroxide, ready to be recycled.

Visions for the future

Road traffic is likely to increase even more. Must road vehicles have to satisfy human prestige, and do we need a strong and heavy vehicle cage to protect drivers and passengers from being injured in case of inevitable mistakes? As soon as wireless, electronic control systems are installed to help to avoid crashes, much lighter and more energy-efficient vehicles will be built in countries that do not consider energy availability to be unlimited. In Europe and Japan, there are already small and lightweight cars available with a combustion engine using as little as 3 l of petrol per 100 km (that is, less than half the amount used by conventional, widely used cars). Four kilograms of hydrogen are enough to run 400 km with a combustion engine, and only 2 kg hydrogen for a modern car driven by a fuel cell.

The new joint venture of Shell, GfE and Hydro-Québec³ on hydrogen storage using metal hydrides, and the fact that no comparable economic effort on hydrogen storage in carbon nanostructures exists, can be taken as clear signs in favour of the metal–hydrogen systems. There is reason for hope that one day much better hydrogen-storage materials will be discovered and developed, rather as we have seen a revolution in high-temperature superconducting materials or hard permanent magnets. We must bear in mind that to develop a sustainable future energy policy requires us to focus not only on the scientific and technical challenge, but also on vital adaptations by the socioeconomic system and a change in attitudes to energy. Sustainability — and humanity — will profit if the price we

pay for energy includes costs for long-term production, transport and distribution of energy, materials, and restoration of the damaged environment. □

- Winter, C. J. & Nitsch, J. *Hydrogen as an Energy Carrier: Technologies, Systems, Economy* (Springer, 1988).
- Bain, A. & Van Vorst, W. D. *Int. J. Hydrogen Energy* **24**, 399–403 (1999).
- Shell Hydrogen, Hydro-Québec (HQ) & Gesellschaft für Elektrometallurgie (GfE). Hydrogen storage joint venture to be established. < <http://www.shell.com> > Press release (12-07-2001).
- Nellis, W. J., Louis, A. A. & Ashcroft, N. W. Metallization of fluid hydrogen. *Phil. Trans. R. Soc. Lond. A* **356**, 119–135 (1998).
- Orimo, S.-I. *et al.* Hydrogen in the mechanically prepared nanostructured graphite. *Appl. Phys. Lett.* **75**, 3093 (1999).
- Orimo, S., Matsuhashima, T., Fujii, H., Fukunaga, T. & Majer, G. Defective carbon for hydrogen storage—thermal desorption property of the mechanically prepared nanostructured graphite. *J. Appl. Phys.* (in the press).
- Stan, G. & Cole, M. W. Hydrogen adsorption in nanotubes. *J. Low Temp. Phys.* **110**, 539–544 (1998).
- Hirscher, M. (ed.) Hydrogen storage in nanoscale carbon and metals. *Appl. Phys. A* (special issue) **72**, 2 (2001).
- Sholl, C. A. & Gray, E. MacA. (eds) Proc. Int. Symp. Metal Hydrogen Systems—Fundamentals and Applications, Noosa, Australia, 1–6 October 2000. *J. Alloys Compounds* (in the press).
- Chambers, A., Park, C., Baker, R. T. K. & Rodriguez, N. M. Hydrogen storage in graphite nanofibers. *Phys. Chem. B* **102**, 4253–4256 (1998).
- Dillon, A. C. *et al.* Storage of hydrogen in single-walled carbon nanotubes. *Nature* **386**, 377–379 (1997).
- Dillon, A. C. *et al.* Carbon nanotube materials for hydrogen storage. Proc. 2000 DOE/NREL Hydrogen program review, 8–10 May 2000.
- Hirscher, M. *et al.* Hydrogen storage in sonicated carbon materials. *Appl. Phys. A* **72**, 129–132 (2001).
- Züttel, A. *et al.* Hydrogen sorption by carbon nanotubes and other carbon nanostructures. *J. Alloys Compounds* (in the press).
- Chen, P., Wu, X., Lin, J. & Tan, K. L. High H₂ uptake by alkali-doped carbon nanotubes under ambient pressure and moderate temperatures. *Science* **285**, 91–93 (1999).
- Hirscher, M. *et al.* Hydrogen storage in carbon nanostructures. *J. Alloys Compounds* (in the press).
- Nijkamp, M. G., Raaymakers, J. E. M. J., Van Dillen, A. J. & De Jong, K. P. Hydrogen storage using physisorption—materials demands. *Appl. Phys. A* **72**, 619–623 (2001).
- Züttel, A. *et al.* Hydrogen storage in carbon nanostructures. *Int. J. Hydrogen Energy* (in the press).
- Enoki, T., Shindo, K. & Sakamoto, N. Electronic properties of alkali-metal-hydrogen-graphite intercalation compounds. *Z. Phys. Chem.* **181**, 75–82 (1993).
- Schlapbach, L. (ed.) *Hydrogen in Intermetallic Compounds I. Electronic, Thermodynamic, and Crystallographic Properties, Preparation* (Topics in Applied Physics Vol. 63) (Springer, 1988).
- Schlapbach, L. (ed.) *Hydrogen in Intermetallic Compounds II. Surface and Dynamic Properties, Applications* (Topics in Applied Physics Vol. 67) (Springer, 1992).
- Sandrock, G. & Thomas, G. The IEA/DOC/SNL on-line hydride databases. *Appl. Phys. A* **72**, 153–155 (2001).
- Sakai, T., Natsukawa, M. & Iwakura, C. Rare earth intermetallics for metal–hydrogen batteries. *Handb. Phys. Chem. Rare Earths* **21**, 135–180 (1995).
- Latroche, M., Percheron-Guegan, A. & Chabre, Y. Influence of cobalt content in $\text{MmNi}_{(1-x)}\text{Mn}_{0.3}\text{Al}_{0.7}\text{Co}$ alloy ($x = 0.36$ and 0.69) on its electrochemical behaviour studied by in situ neutron diffraction. *J. Alloys Compounds* **295**, 637–642 (1999).
- Schlapbach, L., Felix Meli, F., Züttel, A., Westbrook, J. H. & Fleischner, R. L. (eds) in *Intermetallic Compounds: Principles and Practice* Vol. 2, Ch. 22 (Wiley, 1994).
- Zaluska, A., Zaluski, L. & Stroem-Olsen, J. O. Structure, catalysis and atomic reactions on the nanoscale: a systematic approach to metal hydrides for hydrogen storage. *Appl. Phys. A* **72**, 157 (2001).
- Yvon, K. Complex transition metal hydrides. *Chimia* **52**, 613–619 (1998).
- Liu, F. J. & Suda, S. A method for improving the long-term storability of hydriding alloys by air water exposure. *J. Alloys Compounds* **231**, 742–750 (1995).
- Akiba, E. & Iba, H. Hydrogen absorption by Laves phase related BCC solid solution. *Intermetallics* **6**, 461–470 (1998).
- Kuriwa, T. *et al.* New V-based alloys with high protium absorption and desorption capacity. *J. Alloys Compounds* **295**, 433–436 (1999).
- Tsukahara, M. *et al.* Hydrogen storage and electrode properties of V-based solid solution type alloys prepared by a thermic process. *J. Electrochem. Soc.* **147**, 2941–2944 (2000).
- Inoue, H. *et al.* Effect of ball-milling with Ni and Raney Ni on surface structural characteristics of TiV₂1Ni_{0.3} alloy. *J. Alloys Compounds* **325**, 299–303 (2001).
- Bogdanovic, B. & Schwickardi, M. Ti-doped alkali metal aluminium hydrides as potential novel reversible hydrogen storage materials. *J. Alloys Compounds* **253**, 1–9 (1997).

Acknowledgements

We thank the Swiss Federal Office of Energy (BFE), in contract with IEA, the Swiss Federal Office of Education and Science (BBW), and the University of Fribourg and EMPA for support of our hydrogen-storage research projects.

Issues and challenges facing rechargeable lithium batteries

J.-M. Tarascon* & M. Armand†

*Université de Picardie Jules Verne, Laboratoire de Réactivité et Chimie des Solides, UMR-6007, 33 rue Saint Leu, 80039, Amiens, France

†Department of Chemistry, University of Montreal, C.P. 6128 Succ. Centre Ville, Montréal, Quebec H3C 3J7, Canada

Technological improvements in rechargeable solid-state batteries are being driven by an ever-increasing demand for portable electronic devices. Lithium-ion batteries are the systems of choice, offering high energy density, flexible and lightweight design, and longer lifespan than comparable battery technologies. We present a brief historical review of the development of lithium-based rechargeable batteries, highlight ongoing research strategies, and discuss the challenges that remain regarding the synthesis, characterization, electrochemical performance and safety of these systems

Rechargeable Li-ion cells are key components of the portable, entertainment, computing and telecommunication equipment required by today's information-rich, mobile society. Despite the impressive growth in sales of batteries worldwide, the science underlying battery technology is often criticized for its slow advancement. This is true whatever the technology considered (for example, nickel–cadmium, nickel–metal hydride or Li ion). Certainly, when compared, energy storage cannot keep pace with the rate of progress in the computer industry (Moore's law predicts a doubling of memory capacity every two years), yet the past decade has produced spectacular advances in chemistry and engineering within the emerging technologies of Ni–MeH and Li-ion batteries. These cells are now supplanting the well known Ni–Cd batteries.

A battery is composed of several electrochemical cells that are connected in series and/or in parallel to provide the required voltage and capacity, respectively. Each cell consists of a positive and a negative electrode (both sources of chemical reactions) separated by an electrolyte solution containing dissociated salts, which enable ion transfer between the two electrodes. Once these electrodes are connected externally, the chemical reactions proceed in tandem at both electrodes, thereby liberating electrons and enabling the current to be tapped by the user. The amount of electrical energy, expressed either per unit of weight (W h kg^{-1}) or per unit of volume (W h l^{-1}), that a battery is able to deliver is a function of the cell potential (V) and capacity (A h kg^{-1}), both of which are linked directly to the chemistry of the system. Among the various existing technologies (Fig. 1), Li-based batteries — because of their high energy density and design flexibility — currently outperform other systems, accounting for 63% of worldwide sales values in portable batteries¹. This explains why they receive most attention at both fundamental and applied levels.

Historical developments in Li-battery research

Before reviewing the present status of research and future challenges for Li-battery technologies, we present a brief historical account of developments over the past 30 years, as personally perceived.

The motivation for using a battery technology based on Li metal as anode relied initially on the fact that Li is the most electropositive (-3.04 V versus standard hydrogen electrode)

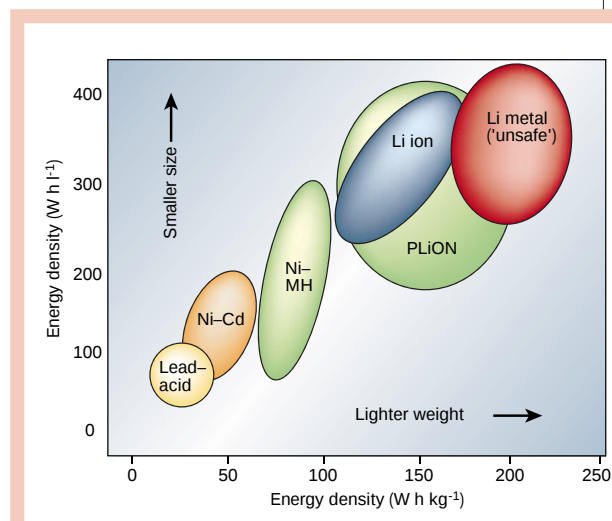


Figure 1 Comparison of the different battery technologies in terms of volumetric and gravimetric energy density. The share of worldwide sales for Ni–Cd, Ni–MeH and Li-ion portable batteries is 23, 14 and 63%, respectively. The use of Pb–acid batteries is restricted mainly to SLI (starting, lighting, ignition) in automobiles or standby applications, whereas Ni–Cd batteries remain the most suitable technologies for high-power applications (for example, power tools).

as well as the lightest (equivalent weight $M = 6.94 \text{ g mol}^{-1}$, and specific gravity $\rho = 0.53 \text{ g cm}^{-3}$) metal, thus facilitating the design of storage systems with high energy density. The advantage in using Li metal was first demonstrated in the 1970s with the assembly of primary (for example, non-rechargeable) Li cells². Owing to their high capacity and variable discharge rate, they rapidly found applications as power sources for watches, calculators or for implantable medical devices. Over the same period, numerous inorganic compounds were shown to react with alkali metals in a reversible way. The discovery of such materials, which were later identified as intercalation compounds, was crucial in the development of high-energy rechargeable Li systems. Like most innovations, development of the technology resulted from several contributions. By 1972, the concept of electrochemical intercalation and its potential use were clearly defined^{3,4}, although the information was not widely disseminated, being reported only in conference proceedings. Before this time, solid-state chemists had been accumulating

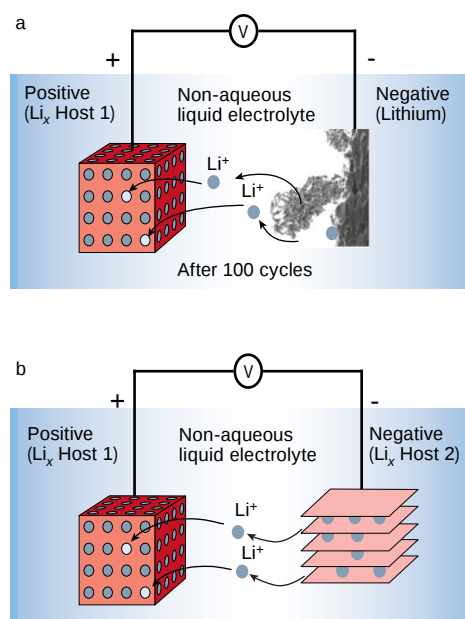


Figure 2 Schematic representation and operating principles of Li batteries.

a. Rechargeable Li-metal battery (the picture of the dendrite growth at the Li surface was obtained directly from *in situ* scanning electron microscopy measurements⁷¹).

b. Rechargeable Li-ion battery.

structural data on the inorganic layered chalcogenides^{5,6}, and merging between the two communities was immediate and fruitful.

In 1972, Exxon^{7,8} embarked on a large project using TiS₂ as the positive electrode, Li metal as the negative electrode and lithium perchlorate in dioxolane as the electrolyte. TiS₂ was the best intercalation compound available at the time, having a very favourable layered-type structure. As the results were published in readily available literature, this work convinced a wider audience. But in spite of the impeccable operation of the positive electrode, the system was not viable. It soon encountered the shortcomings of a Li-metal/liquid electrolyte combination — uneven (dendritic) Li growth as the metal was replated during each subsequent discharge–recharge cycle (Fig. 2a), which led to explosion hazards. Substituting Li metal for an alloy with Al solved the dendrite problem⁹ but, as discussed later, alloy electrodes survived only a limited number of cycles owing to extreme changes in volume during operation. In the meantime, significant advances in intercalation materials had occurred with the realization at Bell Labs that oxides, besides their early interest for the heavier chalcogenides^{10,11}, were giving higher capacities and voltages. Moreover, the previously held belief that only low-dimensional materials

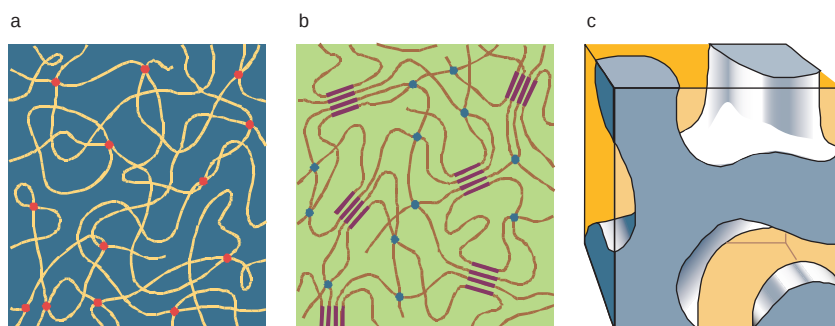
could give sufficient ion diffusion disappeared as a framework structure (V₆O₁₃) proved to function perfectly¹². Later, Goodenough, with Li_xMO₂ (where M is Co, Ni or Mn)^{13,14}, would propose the families of compounds that are still used almost exclusively in today's batteries.

To circumvent the safety issues surrounding the use of Li metal, several alternative approaches were pursued in which either the electrolyte or the negative electrode was modified. The first approach¹⁵ involved substituting metallic Li for a second insertion material (Fig. 2b). The concept was first demonstrated in the laboratory by Murphy *et al.*¹⁶ and then by Scrosati *et al.*¹⁷ and led, at the end of the 1980s and early 1990s, to the so-called Li-ion or rocking-chair technology. The principle of rocking-chair batteries had been used previously in Ni–MeH batteries^{18,19}. Because of the presence of Li in its ionic rather than metallic state, Li-ion cells solve the dendrite problem and are, in principle, inherently safer than Li-metal cells. To compensate for the increase in potential of the negative electrode, high-potential insertion compounds are needed for the positive electrode, and emphasis shifted from the layered-type transition-metal disulphides to layered- or three-dimensional-type transition-metal oxides¹³. Metal oxides are more oxidizing than disulphides (for example, they have a higher insertion potential) owing to the more pronounced ionic character of 'M–O' bonds compared with 'M–S' bonds. Nevertheless, it took almost ten years to implement the Li-ion concept. Delays were attributed to the lack of suitable materials for the negative electrode (either Li alloys or insertion compounds) and the failure of electrolytes to meet — besides safety measures — the costs and performance requirements for a battery technology to succeed. Finally, capitalizing on earlier findings^{20,21}, the discovery of the highly reversible, low-voltage Li intercalation–deintercalation process in carbonaceous material²² (providing that carefully selected electrolytes are used), led to the creation of the C/LiCoO₂ rocking-chair cell commercialized by Sony Corporation in June 1991 (ref. 23). This type of Li-ion cell, having a potential exceeding 3.6 V (three times that of alkaline systems) and gravimetric energy densities as high as 120–150 Wh kg^{–1} (two to three times those of usual Ni–Cd batteries), is found in most of today's high-performance portable electronic devices.

The second approach²⁴ involved replacing the liquid electrolyte by a dry polymer electrolyte (Fig. 3a), leading to the so-called Li solid polymer electrolyte (Li-SPE) batteries. But this technology is restricted to large systems (electric traction or backup power) and not to portable devices, as it requires temperatures up to 80 °C. Shortly after this, several groups tried to develop a Li hybrid polymer electrolyte (Li-HPE) battery²⁵, hoping to benefit from the advantages of polymer electrolyte technology without the hazards associated with the use of Li metal. 'Hybrid' meant that the electrolyte included three components: a polymer matrix (Fig. 3b) swollen with liquid solvent and a salt. Companies such as Valence and Danionics were involved in developing these polymer batteries, but HPE systems never materialized at the industrial scale because Li-metal dendrites were still a safety issue.

With the aim of combining the recent commercial success enjoyed by liquid Li-ion batteries with the manufacturing advantages

Figure 3 Schematic representations of polymer electrolyte networks. **a.** Pure (dry) polymer consisting of entangled chains, through which the Li ions (red points) move assisted by the motion of polymer chains. **b.** A hybrid (gel) network consisting of a semicrystalline polymer, whose amorphous regions are swollen in a liquid electrolyte, while the crystalline regions enhance the mechanical stability. **c.** A poly-olefin membrane (Celgard for instance) in which the liquid electrolyte is held by capillaries.



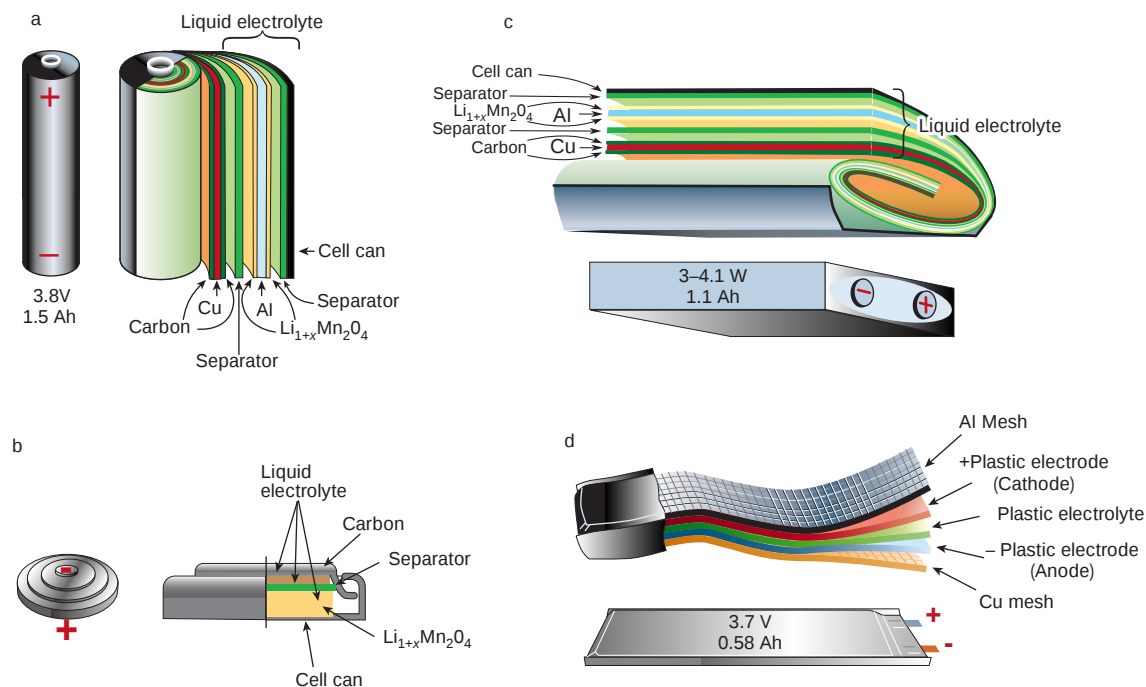


Figure 4 Schematic drawing showing the shape and components of various Li-ion battery configurations. **a**, Cylindrical; **b**, coin; **c**, prismatic; and **d**, thin and flat. Note the

unique flexibility of the thin and flat plastic Li-ion configuration; in contrast to the other configurations, the PLi-ion technology does not contain free electrolyte.

presented by the polymer technology, Bellcore researchers introduced polymeric electrolytes in a liquid Li-ion system²⁶. They developed the first reliable and practical rechargeable Li-ion HPE battery, called plastic Li ion (PLi-ion), which differs considerably from the usual coin-, cylindrical- or prismatic-type cell configurations (Fig. 4). Such a thin-film battery technology, which offers shape versatility, flexibility and lightness, has been developed commercially since 1999, and has many potential advantages in the continuing trend towards electronic miniaturization. Finally, the 'next generation' of bonded liquid-electrolyte Li-ion cells, derived from the plastic Li-ion concept, are beginning to enter the market place. Confusingly called Li-ion polymer batteries, these new cells use a gel-coated, microporous poly-olefin separator bonded to the electrodes (also gel-laden), rather than the P(VDF-HFP)-based membrane (that is, a copolymer of vinylidene difluoride with hexafluoropropylene) used in the plastic Li-ion cells.

Having retraced almost 30 years of scientific venture leading to the development of the rechargeable Li-ion battery, we now describe some of the significant issues and opportunities provided by the field by highlighting the various areas in need of technological advances.

Present status and remaining challenges

Whatever the considered battery technology, measures of its performance (for example, cell potential, capacity or energy density) are related to the intrinsic property of the materials that form the positive and negative electrodes. The cycle-life and lifetime are dependent on the nature of the interfaces between the electrodes and electrolyte, whereas safety is a function of the stability of the electrode materials and interfaces. Compared with mature battery technologies, such as lead-acid or Ni-Cd, rechargeable Li-based battery technologies are still in their infancy, leaving much hope for improvement over the next decade. Such improvements should arise from changes in battery chemistry and cell engineering. Advances in active chemistry are left to the solid-state chemists' creativity and innovation in the design and elaboration of new intercalation electrodes. At the same time, they must bear in mind that it is impossible to predict the demands

that might be placed on tomorrow's portable devices, which in turn places different requirements on the active material chemistry. For instance, with respect to the lower operating voltages of emerging electronics, much debate has focused on whether we should develop a low-voltage active chemistry or rely entirely on electronics (d.c.-d.c. converters) and persist in searching for high-voltage active Li chemistry. Finding the best-performing combination of electrode-electrolyte-electrode can be achieved only through the selective use of existing and new materials as negative and positive electrodes, and of the right electrolyte combination, so as to minimize detrimental reactions associated with the electrode-electrolyte interface — the critical phase of any electrochemical system.

Materials for positive electrodes

The choice of the positive electrode depends on whether we are dealing with rechargeable Li-metal or Li-ion batteries (Fig. 5)²⁷. For rechargeable Li batteries, owing to the use of metallic Li as the negative electrode, the positive electrode does not need to be lithiated before cell assembly. In contrast, for Li-ion batteries, because the carbon negative electrode is empty (no Li), the positive one must act as a source of Li, thus requiring use of air-stable Li-based intercalation compounds to facilitate the cell assembly. Although rechargeable Li-SPE cells mainly use Li-free V_2O_5 or its derivatives as the positive electrode, $LiCoO_2$ is most widely used in commercial Li-ion batteries, deintercalating and intercalating Li around 4 V.

Initially, the use of layered $LiNiO_2$ was considered²⁸, as this displayed favourable specific capacity compared with $LiCoO_2$. But expectations were dismissed for safety reasons after exothermic oxidation of the organic electrolyte with the collapsing delithiated Li_xNiO_2 structure. Delithiated Li_xCoO_2 was found to be more thermally stable than its Li_xNiO_2 counterpart. Thus, substitution of Co for Ni in $LiNi_{1-x}Co_xO_2$ was adopted to provide a partial solution to the safety concerns surrounding $LiNiO_2$.

Although the reversible delithiation of $LiCoO_2$ beyond 0.5 Li is feasible, delithiation for commercial applications has been limited to that value for safety reasons (charged cut-off limited to around

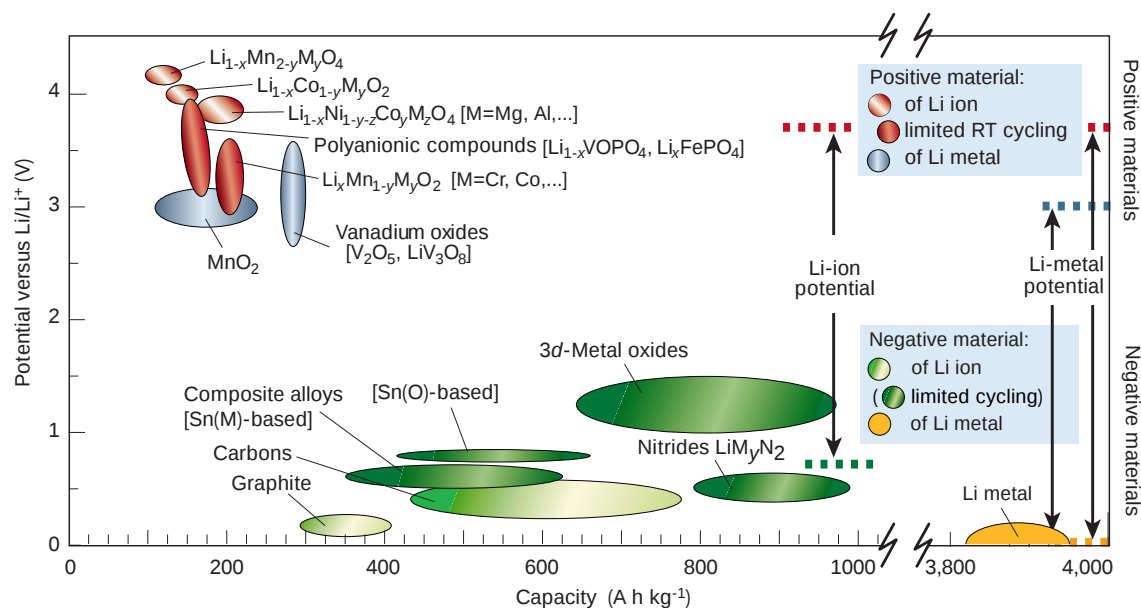


Figure 5 Voltage versus capacity for positive- and negative-electrode materials presently used or under serious considerations for the next generation of rechargeable Li-based cells. The output voltage values for Li-ion cells or Li-metal cells are

represented. Note the huge difference in capacity between Li metal and the other negative electrodes, which is the reason why there is still great interest in solving the problem of dendrite growth.

4.2 V). Several routes were investigated to circumvent these safety and capacity issues. Among them was the successful stabilization of the layered structural framework by an electrochemically inert di-, tri- or tetravalent cationic substitute for Ni or Co (Al, Ga, Mg or Ti). This led to $\text{LiNi}_{1-x}\text{Ti}_{x/2}\text{Mg}_{x/2}\text{O}_2$ phases²⁹, which were claimed to be safe and which displayed practical capacities of 180 mA h g^{-1} compared to only 140 mA h g^{-1} for LiCoO_2 . Another line of investigation involved the synthesis by *chimie douce* ('soft chemistry') of the layered LiFeO_2 and LiMnO_2 phases to take advantage of the $\text{Fe}^{4+}/\text{Fe}^{3+}$ and $\text{Mn}^{4+}/\text{Mn}^{3+}$ redox couples, respectively. In spite of the numerous and diverse synthesis methods, attempts to prepare electrochemically attractive LiFeO_2 phases failed. In contrast, research on LiMnO_2 has been more fruitful³⁰, and the structural instability of the layered phase reversing to the spinel $\text{Li}_x\text{Mn}_2\text{O}_4$ upon cycling has recently been diminished through cationic substitution by chromium ($\text{Li}_{1+x}\text{Mn}_{0.5}\text{Cr}_{0.5}\text{O}_2$)³¹. These materials exhibit a capacity of 190 mA h g^{-1} (larger than that expected from the full oxidation of Mn^{3+} to Mn^{4+}) with little capacity fading upon cycling. It seems that within these materials, the role of Mn is to stabilize the layered structure of the chromium oxide, and that the large capacity is nested in the Cr oxidation state that changes reversibly from +3 to +6. It is therefore unfortunate that Cr presents major toxicity and pricing issues.

The spinel LiMn_2O_4 , although possessing $\approx 10\%$ less capacity than LiCoO_2 , has an advantage in terms of cost and is perceived as being 'green' (that is, non-toxic and from abundant material source). Additionally, it has long been recognized as a potential alternative cathode¹⁴. Its implementation has been delayed because of limited cycling and storage performances at elevated temperatures, although these hurdles were overcome recently by synthesizing dually substituted $\text{LiMn}_{2-x}\text{Al}_x\text{O}_{4-y}\text{F}_y$ spinel phases³², and by altering their surface chemistry³³.

In the search for improved materials for positive electrodes, it has been recognized recently that NaSICON (a family of Na super-ionic conductors) or olivine (magnesium iron silicate) oxyanion scaffolded structures (Fig. 6), built from corner-sharing MO_6 octahedra (where M is Fe, Ti, V or Nb) and XO_4^{n-} tetrahedral anions (where X is S, P, As, Mo or W), offer interesting possibilities³⁴. Polyoxyanionic structures

possess M–O–X bonds; altering the nature of X will change (through an inductive effect) the ionic-covalent character of the M–O bonding. In this way it is possible to systematically map and tune transition-metal redox potentials. For instance, with the use of the phosphate polyanions PO_4^{3-} , the $\text{Fe}^{3+}/\text{Fe}^{2+}$ and $\text{V}^{4+}/\text{V}^{3+}$ redox couples lie at higher potentials than in the oxide form. One of the main drawbacks with using these materials is their poor electronic conductivity, and this limitation had to be overcome^{35,36} through various materials processing approaches, including the use of carbon coatings, mechanical grinding or mixing, and low-temperature synthesis routes to obtain tailored particles. LiFePO_4 , for example, can presently be used at 90% of its theoretical capacity (165 mA h g^{-1}) with decent rate capabilities, and thus is a serious candidate for the next generation of Li-ion cells (Fig. 7). As expected in the light of these promising results, polyoxyanionic-type structures having XO_4^{n-} entities (where X is Si, Ge) are now receiving renewed attention with respect to their electrochemical performance as electrode materials.

Although numerous classes of insertion–deinsertion materials were synthesized over the past 20 years, no real gain in capacity was achieved. One possible way to achieve higher capacities is to design materials in which the metal–redox oxidation state can change reversibly by two units ($\text{M}^{n+2}/\text{M}^n$), while preserving the framework structure, and having molecular masses similar to those of the presently used 3d metal-layered oxides (for example, LiCoO_2). Such an approach is feasible with W-, Mo- or Nb-based metal oxides³⁷, but there is no overall gain in specific energy with these heavier elements. Inserting more than one Li ion per transition metal is also feasible with a few V-based oxides (V^{5+} is reduced to an average state of 3.5 in ' ω - $\text{Li}_3\text{V}_2\text{O}_5$ ' (ref. 38) or to 3.67 in $\text{Li}_5\text{V}_3\text{O}_8$). In principle, except for coordination number requirements, there is no obvious reason why this should not happen with other early transition metals and, in this respect, the recent finding of the reversible $\text{Cr}^{6+}/\text{Cr}^{3+}$ redox couple in a 3d metal-layered compound provides encouragement.

Tuning the morphology or texture of the electrode material to obtain porous and high-surface-area composite electrodes constitutes another exciting, although less exploited, route to enhance electrode capacities³⁹. Indeed, V_2O_5 aerogels, which are

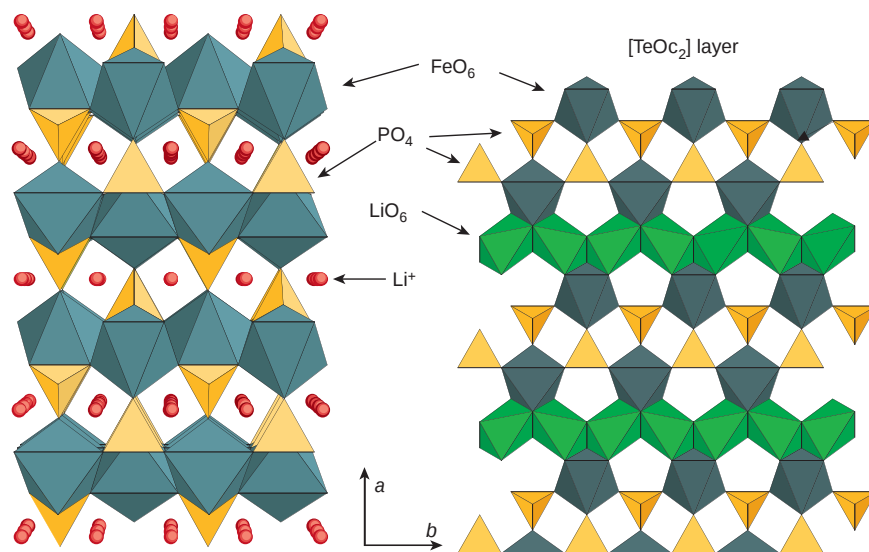


Figure 6 The crystal structure of olivine LiFePO_4 in projection along $[001]$. On the left, expanded view of the framework built on FeO_6 octahedra and PO_4 tetrahedra, with Li ions in red. The FeO_6 octahedra are linked together through corner sharing in the

(b , c) plane. On the right, restricted view of Li, Fe and P distribution between two distorted, h.c.p. (hexagonal close packed) oxygen-dense layers ($\text{P}_{\text{Te}}[\text{LiFe}]_{\text{Oct}}\text{O}_4$). LiO_6 octahedra share edges and Li ions may diffuse along $[010]$ and $[001]$.

mesoporous materials in which nanometre-sized domains are networked through a continuous, highly porous volume of free space, were reported recently to have electroactive capacities up to 100% greater than polycrystalline non-porous V_2O_5 powders and superior power rate capabilities⁴⁰ compared to usual V_2O_5 powders. Such extra capacity apparently derives from the onset of a pure capacitance, associated with the large surface area and high-porosity aerogel matrix, which adds to the existing faradic component. Conductive oxide aerogels such as V_2O_5 and MnO_2 therefore have the potential to boost the field of energy storage once the capacity penalty (in terms of W h l^{-1}) attributable to their poor tap density ($\approx 0.2 \text{ g cm}^{-3}$) is over-

come. Tailor-made nanostructured materials, such as aerogels, create new opportunities not only at the applied level, but also at the fundamental level where some elementary questions, such as the exact mechanism governing these large capacities, remain unanswered.

A radically different approach⁴¹ takes advantage of the facile and reversible redox cleavage of the sulphur–sulphur bond to give lithium thiolate: $-\text{SS}- + 2\text{Li}^+ + 2\text{e}^- \rightleftharpoons -\text{SLi} + \text{LiS}-$. Depending on the electron, the voltage withdrawing power of the moieties attached to the sulphur can be up to 3 V (sulphur itself works at 2.4 V). Although promising in principle in terms of capacity and cost, these systems presently suffer from the relative low density of the reactants and solubility of the resulting thiolates in the electrolyte, leading to self discharge.

Materials for negative electrodes

As a result of numerous chemical (pyrolytic processing) or physical (mechanical milling) modifications, carbon negative electrodes⁴² display electrochemical performances that are improving continuously. Reversible capacities of around 450 mA h g^{-1} are now being reached, compared with a practical value of 350 mA h g^{-1} for graphite (372 mA h g^{-1} for the end compound LiC_6). In parallel, ongoing research efforts are focused on searching for carbon alternatives in the hope of finding materials (Fig. 4) with both larger capacities and slightly more positive intercalation voltages compared to Li/Li^+ , so as to minimize any risks of high-surface-area Li plating at the end of fast recharge, which are associated with safety problems. Such an effort resulted in the emergence of Li transition-metal nitrides as a new potential class of anode materials⁴³, owing to the large, stable and reversible capacity (600 mA h g^{-1}) displayed by one family member, $\text{Li}_{3-x}\text{Co}_x\text{N}$. This result triggered worldwide interest, although performances of the other newly reported Li-based nitrides unfortunately display inferior electrochemical performances compared to the Co phase. Furthermore, use of $\text{Li}_{3-x}\text{Co}_x\text{N}$ is constrained by the restrictive manufacturing requirements for handling such moisture-sensitive negative electrodes.

Throughout the search for carbon alternatives, much effort has been devoted to the use of Li alloys. The first commercial cell was

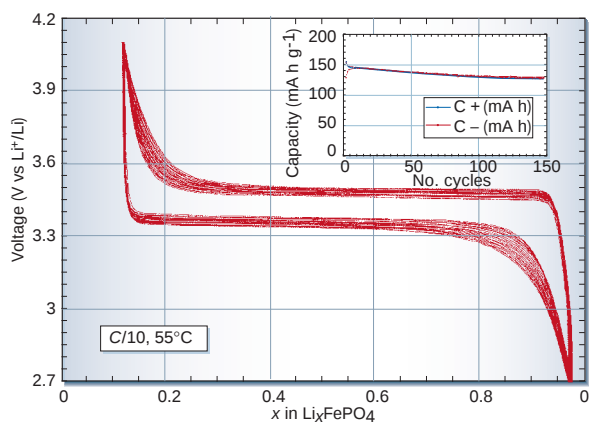


Figure 7 Cycling behaviour at 55°C of an optimized LiFePO_4/C composite electrode (83% of active material) at a scan rate of $\text{C}/10$. Fine particles of LiFePO_4 were obtained from annealing at 500°C a solid intimate mixture resulting from evaporation of an aqueous solution containing Li, Fe(III) and P precursors. The composite electrode was obtained from ball-milling LiFePO_4 with carbon SP. From ref. 36.

introduced in the 1980s by Matsushita; this was based on Wood's metal (a low-melting alloy of Bi, Pb, Sn and Cd), whose cycling performances were found to deteriorate with increased depth of discharge. While attractive in terms of gravimetric capacity, Li alloys suffer from cyclability issues resulting from large Li-driven volume swings (up to 200%), which cause disintegration and hence a loss of electrical contacts between particles. Although a reduction in alloy particle size clearly benefits the cyclability by increasing tolerance to stress cracking, so far the gains are not sufficient⁴⁴. However, it became clear that any physical or chemical means of overcoming the problem of reactant expansion should be beneficial, hence the use of composite negative electrodes. The basis behind this concept is the use of a 'buffer matrix' to compensate for the expansion of the reactants, so preserving the electrical pathway⁴⁵. Initially, such a buffer action was achieved by mixing two alloys that reacted at different potentials so that the electrochemically active phase was imbedded in a non-electrochemically active matrix.

A similar approach held considerable promise in 1997, when Fuji announced the commercialization of a new Li-ion technology (STALION) using an amorphous tin composite oxide (ATCO) as negative electrode. This reacts reversibly with Li at about 0.5 V, and has a specific capacity twice that of graphite⁴⁶. *In situ* X-ray diffraction studies of the ATCO electrode led to the conclusion that the Li reactivity mechanism in these composites was based on oxide decomposition by Li through an initial irreversible process to form intimately mixed Li_2O and metallic Sn, followed by a Li alloying reaction to form nanodomains of $\text{Li}_{4.4}\text{Sn}$ embedded within the Li_2O matrix⁴⁷. However, the STALION cell was never commercialized, in spite of its announcement at the end of 1998. This was most likely due to poor long-term cyclability, the huge and irreversible capacity loss during the first cycle, which was reported by many groups, and the necessity of finding a convenient source for the two initial Li ions needed for the SnO reduction process.

Besides ATCO, other investigations, such as those pursued by Dahn *et al.*⁴⁸ on the Sn-Fe-C system, have also revealed an appealing low-voltage reversible reactivity in composite materials developed as negative electrodes (in spite of initial irreversibility and short-lived capacities). The best experimental proof of the beneficial aspect of the buffer matrix arises from the ability to obtain several hundred cycles on a composite made by precipitating Sn metal at the grain boundaries of electrochemically inactive SnFe_3C grains⁴⁹. However, the cycling performance was improved at the expense of the electrode electrochemical capacity.

A new approach to alleviate the problems of alloy expansion, proposed by Thackeray *et al.*⁵⁰, involved selecting intermetallic alloys such as Cu_6Sn_5 , InSb and Cu_2Sb that show a strong structural relationship to their lithiated products, Li_2CuSn and Li_3Sb for the Sn and Sb compounds, respectively. InSb and Cu_2Sb electrodes are particularly attractive candidates because they operate through a reversible process of lithium insertion and metal extrusion, with an invariant face-centred-cubic Sb array (that is, this array provides a stable host framework for both the incoming and extruded metal atoms). In the ternary $\text{Li}_x\text{In}_{1-y}\text{Sb}$ system ($0 < x < 3$, $0 < y < 1$), the Sb array expands and contracts isotropically by only 4%, whereas the overall expansion of the electrode is 46% if the extruded In is taken into account. This expansion is considerably more favourable than the expansion of binary systems such as LiAl , which expand by ~200% during the phase transition of Al to LiAl . InSb and Cu_2Sb electrodes provide reversible capacities between 250 and 300 mA h g^{-1} . Despite the new and elegant concept behind the design of these intermetallic electrodes, they still suffer from poor cyclability, particularly upon the initial cycle; nevertheless, the approach deserves further study.

Based on the peculiar behaviour (that is, large capacity at low potential) of the transition-metal vanadates M-V-O , first proposed by Fuji Co.⁵¹ and later studied by several groups, Poizot *et al.*⁵² reinvestigated the reactivity of Li-metal oxide. Surprisingly, they found a Li electrochemical activity for well known oxides, but these

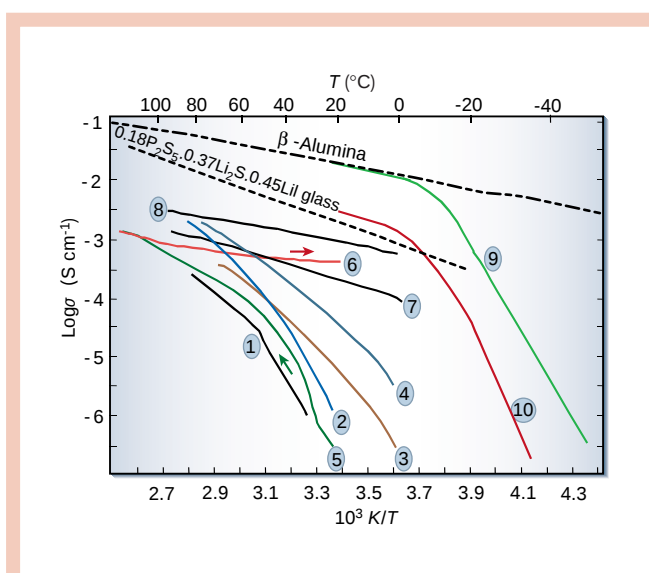


Figure 8 Arrhenius plot of conductivity for various solid electrolytes. 1, First-generation $\text{PEO-LiCF}_3\text{SO}_3$; 2, new solutes with high-dissociation $\text{PEO-Li}[(\text{CF}_3\text{SO}_2)_2\text{N}]$; 3, low- T_g combination polymer; 4, plasticized polymer electrolyte $\text{PEO-Li}[(\text{CF}_3\text{SO}_2)_2\text{N}] + 25\%$ w/w PEG-dimethylether (molecular weight, 250); liquid crystalline polymer electrolytes; 5, heating curves; 6, cooling curve⁶⁴; 7, gel-type polymer (X-linked PEO -dimethacrylate- $\text{Li}[(\text{CF}_3\text{SO}_2)_2\text{N}]\text{-PC}$ 70%); 8, liquid electrolyte $\text{PC/DME LiCF}_3\text{SO}_3$; 9, liquid electrolyte EC/DMC-LiPF_6 at low temperature⁶¹; 10, gel electrolyte $\text{P(VDF-HFP)/EC/DMC-LiPF}_6$ (ref. 61).

did not react with Li according to the classical processes of Li insertion–deinsertion or Li alloying, the catchwords of the past 20 years. For instance, MO-type compounds (where M is Co, Ni, Fe, Cu or Mn), having a rocksalt structure and containing metal elements (M) that do not alloy with Li, exhibited capacities two to three times those of carbon with 100% capacity retention for up to 100 cycles. The mechanism of Li reactivity in such materials differs from the classical processes, and is nested in the electrochemically driven, *in situ* formation of metal nanoparticles during the first discharge, which enables the formation and decomposition of Li_2O upon subsequent cycling⁵³. Remaining issues relate to a problem of surface, with the chemical reactivity being enhanced as the particle size becomes smaller. These findings open new avenues of research aimed at capitalizing on the beneficial effect that particle-size confinement could have within the field of electrochemistry. These and related nanocluster systems under development hold much promise for future developments.

Polymer and liquid electrolytes

Besides the electrodes, the electrolyte, which commonly refers to a solution comprising the salts and solvents, constitutes the third key component of a battery. Although the role of electrolyte is often considered trivial, its choice is actually crucial, and is based on criteria that differ depending on whether we are dealing with polymer or liquid-based Li-ion rechargeable batteries⁵⁴. For instance, working with a highly oxidizing ($>4\text{V}$ versus Li/Li^+) positive electrode material for Li-ion batteries requires electrolyte combinations that operate well outside their window of thermodynamic stability (3.5 V). This is the reason why early workers in the field ignored very positive cathode materials. But fortunately this electrolyte stability is kinetically controlled, enabling the use of non-aqueous electrolytes at potentials as high as 5.5 V (ref. 55). Similarly, the use of a polymer rather than a liquid electrolyte adds further selection criteria linked to the electrochemical stability of the polymer⁵⁶. There are numerous liquid solvents available, each with different dielectric constants and viscosity, and we can select specific solvents to favour the ionic

conductivity of the electrolyte. In contrast, there are only a few Li-based salts or polymers to choose from, the most commonly used ones being based on polyethylene oxide (PEO). The results from research efforts aimed at counterbalancing this deficit have led to the present level of research and development on electrolytes.

Guided by general concepts of viscosity and dielectric constants, optimizing the ionic conductivity of a liquid electrolyte almost becomes a field-trial approach with the hope of finding the key ingredients. For instance, only ethylene carbonate can provide the *ad hoc* protective layer on the surface of graphite that prevents further reaction (continuous electrolyte reduction and self-discharge). Ethylene carbonate is therefore present in almost all commercial compositions, thinned with other solvents owing to its high melting point. Why the homologous propylene carbonate is unsuitable for this protective layer remains an open question, reminding us that chemistry has its secrets.

In contrast, achieving high ionic conductivity in Li-based polymer electrolytes requires a better understanding of the fundamentals of ion dissociation and transport. Both the nature of the polymer–salt interaction and the precise structure of highly concentrated electrolyte solutions have always resisted rationalization. Nevertheless, a principal goal has been to search for new, highly conductive salts with a large electrochemical window, which form a eutectic composition with PEO that melts at the lowest possible temperature⁵⁷. The concept of non-coordinating anions with extensive charge delocalization was achieved with the perfluorosulphonimide $\text{Li}^+[\text{CF}_3\text{SO}_2\text{NSO}_2\text{CF}_3]^-$ salt (abbreviated as LiTFSI)⁵⁸. Figure 8 shows the marked improvement when passing, with simple PEO, from a ‘conventional’ LiCF_3SO_3 salt (curve 1) to the imide salt (curve 2), where an order of magnitude is gained, not ignoring the larger elastomeric domain towards low temperature. The polymer architecture has a role independently of dissociation. Attaching the side chains of the solvating group to the polymer increases the degrees of freedom as a result of dangling chain ends; this improves conductivity (Fig. 8, curve 3), but compromises the mechanical properties.

Although efforts aimed at enhancing the ionic conductivity of polymer electrolytes have been insufficient to permit operation at room temperature, they have benefited liquid-based electrolyte systems in terms of cost and safety, so that battery manufacturers of Li-ion cells are eager to see the further development of organic anion-based salts able to operate at voltages greater than 4.5 V. LiTFSI is an example of this cross-fertilization. Although extremely resistant to oxidation itself, the electrochemical use of such a salt is limited to 4 V in presence of an Al collector, because a stable and soluble Al salt can be formed as a consequence of the robustness of the anion bonds. With the less stable coordination anions (LiPF_6), decomposition occurs immediately and is accompanied by formation of protective AlF_3 . However, owing to its high conductivity in any medium, its safety and lack of toxicity, LiTFSI is being used increasingly in Li-ion batteries, the corrosion problem having been solved by simple addition of a passivating coordination-type salt. A wide range of anion-forming systems now exists, especially in the imide family, and these are viewed as candidates for high conductivity and Al passivation.

Having exploited most of the possibilities offered by ‘dry’ polymers to improve conductivity (ability, amorphous state and lowest possible glass-transition temperature T_g controlling the ion mobility), a remaining option was to use additives, known in polymer science as plasticizers, to act as chain lubricants, so leading to the development of ‘hybrid’ polymer electrolytes⁵⁹. Indeed, suitable plasticizers are chosen between the same polar solvents as for liquid electrolytes⁶⁰, such as propylene carbonate, γ -butyrolactone or polyethylene glycol ethers, or are formed from short-chain PEO (4–25 monomer units). A lightly plasticized material (10–25% additive) improves conductivity by an order of magnitude (Fig. 8, curve 4). Gels, on the other hand, contain 60–95% liquid electrolyte, and are only 2–5 times less conductive than their liquid counterpart⁶¹

(Fig. 8, curves 7–10). Interestingly, when the gelling agent is a polyether, most of the solvation still takes place through the polymer chains rather than the carbonate solvents, the latter being less prone to donate electron pairs. Understandably, the lightly plasticized systems can be used in a Li-metal configuration, as much of the resilience of the pristine polymer is retained, whereas the much softer gels require a Li-ion configuration.

It is surprising that in spite of the direct link between their ionic conductivity and their degree of amorphicity, very little is known about the structural chemistry of polymer electrolytes. In contrast to the well established dynamic view of ionic conductivity on these materials, Bruce *et al.*⁶² recently proposed a structural view, highlighting the importance of aligning or organizing the polymer chains in order to enhance the levels of ionic conductivity. Similarly, Wright and co-workers⁶³ and Ingram⁶⁴ focused on the liquid crystalline state to force the solvating polymer into a conformation that was dictated by the liquid crystal part. The result is a partial decoupling of the conductivity from the glass-transition temperature of the polymer. The conductivity of such liquid crystalline chain polymers is low at room temperature, but reaches liquid-like values at high temperature or when kept under polarization, and remains so upon cooling to room temperature (Fig. 8, curves 5 and 6), without appreciable activation energy⁶⁴. As these new perspectives generate renewed interest in the design of polymer electrolytes, it is hoped that solutions may eventually be found to the problems of ionic conductivity afflicting this class of materials at ambient or subambient temperature.

The addition of nanoparticle fillers (10% w/w), such as Al_2O_3 or TiO_2 , to simple PEO compounds increases the conductivity several-fold at 60–80 °C, and prevents crystallization for at least several weeks at room temperature⁶⁵. Two important advantages of these systems are an increase in the apparent Li transport number, from a low of ≈ 0.3 (common to polymer, liquid and gels) to ≈ 0.6 , and the formation of a stable, low-resistance interface in contact with Li. Because these materials obey different conduction mechanisms, they are presently the focus of many studies, both practical and theoretical⁶⁶.

Technologies based on either solid polymer or ‘hybrid’ polymer electrolytes offer great advantages that will be necessary to meet the flexible, shape-effective requirements dictated by today’s electronic miniaturization, while at the same time providing a larger autonomy. Current Li technologies rely on liquid-jellyroll or prismatic-cell configurations. Neither fits well in a multiple-cell configuration. This is in marked contrast with the recent thin, plate-like plastic (PLiON) technology that enables excellent packing efficiency, as multiple plates can be densely packaged in parallel within one cell while preserving the flexibility of the overall package. Future technology improvements should focus on better chemical engineering of the bonded laminates, so as to obtain even thinner cells. Similar attributes can be provided by the solid Li-polymer technology, which in addition exhibits extra capacity and is free of electrolyte leakage. This currently operates at 80 °C. Although warm temperatures may be an advantage for the large batteries required by the transportation sector, problems of conductivity have to be solved for electronic applications, as emphasized earlier.

The electrode–electrolyte interface

The Li-ion cell density can be improved through a selective use of appropriate existing or new materials for negative and positive electrodes. However, optimizing an electrode material is only the first step in the process leading to its implementation in a practical cell. Indeed, while the capacity of a cell is nested in the structural or electronic behaviour of its electrode, poor cell lifetimes are rooted mainly in side reactions occurring at the electrode–electrolyte interface. Thus, mastering the chemical stability of any new electrode material with respect to its operating liquid or polymer electrolyte medium, which requires a control of the electrode–electrolyte interface through surface chemistry, is as important as designing new

materials. Tackling interfacial issues is both tedious and complex. We should remember that, despite many years of research devoted to the mechanism by which the solid electrolyte interphase forms on Li or carbonaceous materials, its composition and nature are still the subject of much controversy. In contrast, the positive electrode interface has received little attention over the years, despite its equally crucial role. Its importance is amplified with the Li-ion technology, where high voltages exceed the electrochemical resistance of the electrolyte oxidation, and even favour its catalytically driven decomposition. Thus, it is critical to control the electrode surface so as to modify its catalytic activity towards electrolyte decomposition. The strategy developed to address this issue uses coatings that encapsulate, through chemical or physical means, the electrode grains with either an inorganic or an organic phase. This concept, successfully applied to the spinel LiMn_2O_4 , is based on minimizing the surface area of the active material in direct contact with the electrolyte³³. The coating must allow easy diffusion of Li ions and, although insulating in nature, must be thin enough to allow the electrons to tunnel through. Equally relevant is the unexplained role of filler additives in polymer electrolytes⁶⁵, which markedly reduce the interfacial impedance in contact with Li.

Thirty years after its initial observation, the key issue of Li dendrite growth, which was thought to be governed mainly by current densities, remains highly topical, especially in light of recent promising results obtained by Aurbarch's and Bates' groups. Revisiting Exxon's solvent 1-3-dioxolane, Aurbarch and co-workers⁶⁷ showed that the use of LiAsF_6 salt led to a completely different Li morphology from that obtained from an ethylene carbonate–dimethyl carbonate (EC–DMC) electrolyte. They explained this in terms of the reactivity of dioxolane with lithium, which forms an elastomeric coating endowing the Li surface with plasticity and flexibility, thereby reducing dendrite growth. These findings were implemented in Li/ MnO_2 commercial Tadiran cells that, under well defined cycling conditions, are claimed to be safe. The bulk polymerization of the cyclic ether, initiated at the positive electrode on overcharge, acts as a thermal shutdown. Even more spectacular are recent reports by Bates *et al.*⁶⁸ who succeeded in cycling LiCoO_2/Li thin-film batteries for more than 50,000 cycles using a glassy electrolyte in $\approx 1\text{-}\mu\text{m}$ -thick films obtained by sputtering techniques. By controlling the uniform Li stripping–plating mechanism, the same authors demonstrated the feasibility of a Li-free, rechargeable, thin-film battery — that is, cells constructed in the discharged state with no Li metal initially present⁶⁹. Such findings, whether resulting from low-current density or the use of solid electrolyte, show that the problem of dendrite growth can be solved, at least with special cell configurations. Visco and co-workers⁷⁰ recently showed that a glassy nanometric layer deposited on Li metal completely insulates it from its environment, even in the presence of liquids, and that this coating can be applied at a high production rate. With further work devoted to the implementation of these findings to large-size Li batteries, the development of a Li-free rechargeable battery remains a realistic goal for the future.

The principal challenge for Li-based rechargeable batteries, or indeed for any battery, lies in gaining better understanding and control of the electrode–electrolyte interface in the hope of designing new solid–solid or solid–liquid interfaces. For example, the nature of the secondary reactions occurring at high temperature, which cause cell failure, remains an unanswered question that must be addressed to ensure the practical success of these technologies. In this case, however, the main difficulty stems from a lack of available techniques to probe the evolution of the electrode–electrolyte interface at a local level. We have so far relied (with the exception of X-ray diffraction) on post-mortem rather than *in situ* studies to determine how the electrodes or interfaces age with time either under cycling or storage conditions, thereby missing key information. But introduction of the plastic Li-ion-type technology has created new opportunities to perform a wide variety of *in situ* characterization techniques. These include X-ray absorption near-edge structure, nuclear magnetic

resonance and Mössbauer spectroscopies, or even scanning electron microscopy observations that allow real-time visualization of dendrite growth at an interface⁷¹. Efforts aimed at developing new characterization tools must be vigorously pursued so as to create a comprehensive database on the electrode–electrolyte interface.

Conclusion

Consumers are in constant demand for thinner, lighter, space-effective and shape-flexible batteries with larger autonomy. Such demand will continue to generate much research activity towards the development of new cell configurations and new chemistries. In this review we hope to have conveyed the message that the field of energy storage is advancing faster than it perhaps has ever done in the past. The benefits, in terms of weight, size and design flexibility provided by today's state-of-the-art Li-ion configurations, which owe much to the design engineers' striving to develop efficient, economical microtechnologies, are a good illustration. The Li-based battery chemistry is relatively young, and as such is a source of aspirations as well as numerous exciting challenges. The latter are not limited to solid-state chemists. The effort should be highly multidisciplinary with strong roots in the fields of organic and inorganic chemistry, physics, surface science and corrosion. Through materials design we can expect significant improvements in energy density. And although designing new materials can be intuitive or based on chemical concepts, coupling these efforts with those of theorists who are able to perform band-structure calculations on envisioned compounds will prove to be highly beneficial. Of equal importance is a better understanding of the electrode–electrolyte interface to facilitate design of new interfaces. Here the goal is well defined, although we must diverge from the empirical approach used so far, and make full use of the recent progress achieved by *in situ* characterization. As Li-rechargeable batteries enter their teenage years, scientists and engineers predict an even brighter future lies ahead. □

1. Takeshita, H. Portable Li-ion, worldwide. Proc. Conf. Power 2000, San Diego, 25 September 2000.
2. Ikeda, H., Saito, T. & Tamura, H. in *Proc. Manganese Dioxide Symp.* Vol. 1 (eds Kozawa, A. & Brodd, R. H.) (IC sample Office, Cleveland, OH, 1975).
3. Steele, B. C. H. in *Fast Ion Transport in Solids* (ed. Van Gool, W.) 103–109 (North-Holland Amsterdam, 1973).
4. Armand, M. B. in *Fast Ion Transport in Solids* (ed. Van Gool, W.) 665–673 (North-Holland Amsterdam, 1973).
5. Rouxel, J., Danot, M., & Bichon, M. Les composites intercalaires Na_xTiS_2 . Etude générale des phases Na_xTiS_2 et K_xTiS_2 . *Bull. Soc. Chim.* **11**, 3930–3936 (1971).
6. Di Salvo, F. J., Schwall, R., Geballe, T. H., Gamble, F. R. & Osieki, J. H. Precursor effects of superconductivity up to 35 K in layered compounds. *Phys. Rev. Lett.* **27**, 310–313 (1971).
7. Whittingham, M. S. Electrochemical energy storage and intercalation chemistry. *Science* **192**, 1226 (1976).
8. Whittingham, M. S. Chalcogenide battery. US Patent 4009052.
9. Rao, B. M. L., Francis, R. W. & Christopher, H. A. Lithium–aluminium electrodes. *J. Electrochem. Soc.* **124**, 1490–1492 (1977).
10. Broadhead, J. & Butherus, A. D. Rechargeable non-aqueous battery. US Patent 3791867.
11. Broadhead, J., DiSalvo, F. J. & Trumbore, F. A. Non-aqueous battery using chalcogenide electrode. US Patent 3864167.
12. Murphy, D. W. & Christian, P. A. Solid state electrodes for high energy batteries. *Science* **205**, 651–656 (1979).
13. Mizushima, K., Jones, P. C., Wiseman, P. J. & Goodenough, J. B. LiCoO_2 ($0 < x \leq 1$): a new cathode material for batteries of high energy density. *Mat. Res. Bull.* **15**, 783–789 (1980).
14. Thackeray, M. M., David, W. I. F., Bruce, P. G. & Goodenough, J. B. Lithium insertion into manganese spinels. *Mat. Res. Bull.* **18**, 461–472 (1983).
15. Armand, M. B. in *Materials for Advanced Batteries* (Proc. NATO Symp. Materials Adv. Batteries) (eds Murphy, D. W., Broadhead, J. & Steele, B. C. H.) 145–161 (Plenum, New York, 1980).
16. Murphy, D. W., DiSalvo, F. J., Carides, J. N. & Waszczak, J. V. Topochemical reactions of rutile related structures with lithium. *Mat. Res. Bull.* **13**, 1395–1402 (1978).
17. Lazzari, M. & Scrosati, B. A cyclable lithium organic electrolyte cell based on two intercalation electrodes. *J. Electrochem. Soc.* **127**, 773–774 (1980).
18. Will, F. G. Hermetically sealed secondary battery with lanthanum nickel anode. US patent 3874958 (1975).
19. Percheron-Guegan, A., Achard, J. C., Sarradin, J. & Bronoël, G. Alliages à base de Lanthane et de Nickel et leurs applications électrochimiques. French patent 7516160 (1975).
20. Guérard, D. & Hérol, A. New method for the preparation of lithium insertion compounds in graphite. *C.R. Acad. Sci. C* **275**, 571–572 (1972).
21. Basu, S. Ambient temperature rechargeable battery. US patent 4,423,125 (filing date, 13 September 1982; publication date, 27 Dec 1983).
22. Mohri, M. *et al.* Rechargeable lithium battery based on pyrolytic carbon as a negative electrode. *J. Power Sources* **26**, 545–551 (1989).
23. Nagaura, T. & Tozawa, K. Lithium ion rechargeable battery. *Prog. Batteries Solar Cells* **9**, 209 (1990).

24. Armand, M., Chabagno, J. M. & Duclot, M. J. in *Fast Ion Transport in Solids Electrodes and Electrolytes* (eds Vashishta, P., Mundy, J.-N. & Shenoy, G. K.) 131–136 (North-Holland, Amsterdam, 1979).
25. Kelly, I. E., Owen, J. R. & Steel, B. H. Poly(ethyleneoxide) electrolytes for operation at near room temperature. *J. Power Sources* **14**, 13–21 (1985); *Interfacial Electrochem.* **168**, 467 (1984).
26. Tarascon, J.-M., Gozdz, A. S., Schmutz, C., Shokoohi, F. & Warren, P. C. Performance of Bellcore's plastic rechargeable Li-ion batteries. *Solid State Ionics* **86–88**, 49–54 (1996).
27. Guyomard, D. in *New Trends in Electrochemical Technology: Energy Storage Systems for Electronics* Vol. 9 (eds Osaka, T. & Datta, M.) 253–350 (Gordon & Breach Science Publishers, 2000).
28. Dahn, J. R., Von Sacken, U., Jozkow, M. W. & Al-Janaby, H. Rechargeable LiNiO₂/carbon cells. *J. Electrochem. Soc.* **138**, 2207–2211 (1991).
29. Yuan Gao, Yakovleva, M. V. & Ebner, W. B. Novel LiNi_{1-x}Ti_xMg_{0.2}O₂ compounds as cathode materials for safer lithium-ion batteries. *Electrochem. Solid State Lett.* **1**, 117–119 (1998).
30. Armstrong, A. R. & Bruce, P. G. Synthesis of layered LiMnO₂ as an electrode for rechargeable lithium batteries. *Nature* **381**, 499–500 (1996).
31. Ammundsen, B. *et al.* in *Proc. Int. Symp. Electrochem. Soc.* Vol. 99-24, 57–67 (ECS, Pennington, NJ, 2000).
32. Amatucci, G. G., Pereira, N., Zheng, T. & Tarascon, J.-M. Failure mechanism and improvement of the elevated temperature cycling of LiMn₂O₄ compounds through the use of the LiAl_{0.5}Mn_{1.5}O_{4-x}F_x solid solution. *J. Electrochem. Soc.* **148**, A171–A182 (2001).
33. Amatucci, G. G., du Pasquier, A., Bly, A., Zheng, T. & Tarascon, J.-M. The elevated temperature performance of the LiMn₂O₄/C system: failure and solutions. *Electrochem. Acta* **45**, 255–271 (1999).
34. Padhi, A. K., Nanjundaswamy, K. S., Masquelier, C., Okada S. & Goodenough, J. B. Effect of structure on the Fe³⁺/Fe²⁺ redox couple in iron phosphates. *J. Electrochem. Soc.* **144**, 1609–1613 (1997).
35. Ravet, N. *et al.* Improved iron-based cathode material. Abstr. No. 127, ECS Fall meeting, Hawaii, 1999.
36. Morcrette, M., Wurm, C., Gaubicher, J. & Masquelier, C. Polyanionic structures as alternative materials for lithium batteries. Abstr. No. 93, Electrode Materials Meeting, Bordeaux Arcachon, 27 May–1 June 2001.
37. Cava, R. J., Murphy, D. W. & Zahurak, S. M. Lithium insertion in Wadsley-Roth phases based on Niobium oxide. *J. Electrochem. Soc.* **30**, 2345–2351 (1983).
38. Delmas, C., Brethes, S. & Menetrier, M. Li₃V₂O_{5-ω}, un nouveau matériau d'électrode pour accumulateur au lithium. *CR Acad. Sci.* **310**, 1425–1430 (1990).
39. Le, D. B. *et al.* High surface area V₂O₅ aerogel intercalation electrodes. *J. Electrochem. Soc.* **143**, 2099–2104 (1996).
40. Dong, W., Rolison, D. R. & Dunn, B. Electrochemical properties of high surface area vanadium oxides aerogels. *Electrochem. Solid State Lett.* **3**, 457–459 (2000).
41. Visco, S. J. & de Jonghe, L. C. in *Handbook of Solid-State Batteries and Capacitors* (ed. Munshi, M. Z. A.) 515 (World Scientific, Singapore, 1995).
42. Dahn, J. R. *et al.* Carbon and graphites as substitutes for the lithium anode. *Industrial Chemistry Library* Vol. 5 (ed. Pistoia, G.) (1994).
43. Shodai, T., Okada, S., Tobishima, S. & Yamabi, I. Study of Li_{3-x}M₂N (M=Co, Ni or Cu) system for use as anode in lithium rechargeable cells. *Solid State Ionics* **86–88**, 785–789 (1996).
44. Winter, M. & Besenhard, J. O. Electrochemical lithiation of tin and tin-based intermetallics and composites. *Electrochem. Acta* **45**, 31–50 (1999).
45. Anani, A., Crouch-Baker, S. & Huggins, R. A. Kinetics and thermodynamic parameters of several binary alloys negative electrode materials at ambient temperature. *J. Electrochem. Soc.* **134**, 3098–3102 (1987).
46. Idota, Y., Kubota, T., Matsufuji, A., Maekawa, Y. & Miyasaka, T. Tin-based amorphous oxide: a high capacity lithium-ion storage material. *Science* **276**, 1395–1397 (1997).
47. Courtney, I. A. & Dahn, J. R. Electrochemical and in situ X-ray diffraction studies of the reaction of lithium with tin oxide composites. *J. Electrochem. Soc.* **144**, 2045–2052 (1997).
48. Mao, O., Dunlap, R. A. & Dahn, J. R. Mechanically alloyed Sn-Fe(-C) powders as anode materials for Li-ion batteries. I. The Sn₂Fe-C system. *J. Electrochem. Soc.* **146**, 405–413 (1999).
49. Beaulieu, L. Y., Larcher, B., Dunlap, R. A. & Dahn, J. R. Reaction of Li with grain-boundary atoms in nano structured compounds. *J. Electrochem. Soc.* **147**, 3206–3212 (2000).
50. Kepler, K. D., Vaughey, J. T. & Thackeray, M. M. Li₂Cu₂Sn₃ (0<x<13): an intermetallic insertion electrode for rechargeable lithium batteries. *Electrochem. Solid State Lett.* **2**, 307 (1999).
51. Idota, Y. *et al.* Nonaqueous battery. US Patent No. 5,478,671 (1995).
52. Poizat, P., Laruelle, S., Grugeon, S., Dupont, L. & Tarascon, J.-M. Nano-sized transition-metal oxides as negative-electrode material for lithium-ion batteries. *Nature* **407**, 496–499 (2000).
53. Poizat, P., Laruelle, S., Grugeon, S., Dupont, L. & Tarascon, J.-M. Electrochemical reactivity and reversibility of cobalt oxides towards lithium. *C.R. Acad. Sci. II* 681–691 (2000).
54. Dominey, L. A. Current state of the art on lithium battery electrolytes. *Industrial Chemistry Library* Vol. 5 (ed. Pistoia, G.) 137–165 (1994).
55. Guyomard, D. & Tarascon, J. M. High voltage stable liquid electrolytes for Li_{1-x}Mn₂O₄/carbon rocking-chair lithium batteries. *J. Power Sources* **54**, 92–98 (1995).
56. Armand, M. The history of polymer electrolytes. *Solid State Ionics* **69**, 309–319 (1994).
57. Fenton, D. E., Parker, J. M. & Wright, P. V. Complexes of alkali metal ions with poly(ethylene oxide). *Polymer* **14**, 589 (1973).
58. Armand, M., Gorecki, W. & Andreani, R. in *Second International Meeting on Polymer Electrolytes* (ed. Scrosati, B.) 91–96 (Elsevier, London, 1989).
59. Fauteux, D. in *Polymer Electrolytes Reviews II* (eds MacCallum, J. R. & Vincent, C. A.) 212 (Elsevier, London, 1989).
60. Feuillade, G. & Perche, P. Ion conductive macromolecular gels and membranes for solid lithium cells. *J. Appl. Electrochem.* **5**, 63–69 (1975).
61. Stallworth, P. E. *et al.* NMR, DSC and high pressure electrical conductivity studies of liquid and hybrid electrolytes. *J. Power Sources* **81**, 739–747 (1999).
62. MacGlashan, G. S., Andreev, Y. G. & Bruce, P. Structure of the polymer electrolyte poly(ethylene oxide): LiAsF₆. *Nature* **398**, 792–793 (1999).
63. Zheng, Y., Chia, F., Ungar, G. & Wright, P. V. Self-tracking in solvent-free, low-dimensional polymer electrolyte blends with lithium salts giving high ambient DC conductivity. *Chem. Commun.* **16**, 1459–1460 (2000).
64. Imrie, C. T., Ingram, M. D. & McHattie, G. S. Ion transport in glassy polymer electrolytes. *J. Phys. Chem. B* **103**, 4132–4138 (1999).
65. Croce, F., Appetecchi, G. B., Persi, L. & Scrosati, B. Nanocomposite polymer electrolytes for lithium batteries. *Nature* **394**, 456–458 (1998).
66. Sata, N., Eberman, K., Eberl, K. & Maier, J. Mesoscopic fast ion conduction in nanometre-scale planar heterostructures. *Nature* **408**, 946–949 (2000).
67. Aurbach, D. Review of selected electrode solution interactions which determine the performance of Li and Li-ion batteries. *J. Power Sources* **89**, 206–218 (2000).
68. Bates, J. B., Dudney, N. J., Neudecker, B., Ueda, A. & Evans, C. D. Thin film lithium and lithium-ion batteries. *Solid State Ionics* **135**, 33–45 (2000).
69. Neudecker, B. J., Dudney, N. J. & Bates, J. B. Lithium-free thin film battery with in situ plated Li anode. *J. Electrochem. Soc.* **147**, 517–523 (2000).
70. Visco, S. J. The development of reversible lithium metal electrodes for advances Li/S batteries. Conf. Proc. Int. Meeting on Power Sources for Consumer and Industrial Applications, Hawaii, 3–6 September 2001.
71. Orsini, F. *et al.* In situ SEM study of the interfaces in plastic lithium cells. *J. Power Sources* **81–82**, 918–921 (1999).

Acknowledgements

The authors thank their colleagues, both in academic institutions and industry, for sharing the gratifying dedication to this field of progress, and P. Rickman for help drawing the figures.

High- T_c superconducting materials for electric power applications

David Larbalestier, Alex Gurevich, D. Matthew Feldmann & Anatoly Polyanskii

Applied Superconductivity Center, Department of Materials Science and Engineering, Department of Physics, University of Wisconsin, Madison, Wisconsin 53706 USA (e-mail: larbalestier@engr.wisc.edu)

Large-scale superconducting electric devices for power industry depend critically on wires with high critical current densities at temperatures where cryogenic losses are tolerable. This restricts choice to two high-temperature cuprate superconductors, $(\text{Bi,Pb})_2\text{Sr}_2\text{Ca}_2\text{Cu}_3\text{O}_x$ and $\text{YBa}_2\text{Cu}_3\text{O}_x$, and possibly to MgB_2 , recently discovered to superconduct at 39 K. Crystal structure and material anisotropy place fundamental restrictions on their properties, especially in polycrystalline form. So far, power applications have followed a largely empirical, twin-track approach of conductor development and construction of prototype devices. The feasibility of superconducting power cables, magnetic energy-storage devices, transformers, fault current limiters and motors, largely using $(\text{Bi,Pb})_2\text{Sr}_2\text{Ca}_2\text{Cu}_3\text{O}_x$ conductor, is proven. Widespread applications now depend significantly on cost-effective resolution of fundamental materials and fabrication issues, which control the production of low-cost, high-performance conductors of these remarkable compounds.

The potential applications of superconductors — Kamerlingh Onnes's name for materials that lose electric resistance on cooling below a transition temperature, T_c — were apparent to Onnes almost immediately. In 1913, just two years after his discovery, Onnes talked in Chicago about the design of very powerful magnets far exceeding the fields achievable by iron; these would cost as much as a battleship if made from copper and cooled with liquid air, but be affordable if made from superconducting wires. By that time he had already tested a Ni alloy coated with Pb-rich superconducting solder, but this lost superconductivity at fields of less than 50 mT. He ascribed this unexpected setback to bad places in the wire, a problem he anticipated would soon be fixed without difficulty!

In fact, applications had to wait 50 more years, particularly because the physics of superconductivity in magnetic fields were seriously misunderstood¹. This need not have been so, because London and Shubnikov made important breakthroughs in understanding the magnetic properties of superconductors in the 1930s. By careful alloying experiments, Shubnikov pointed out the vital distinction between type I superconductors, in which currents flow only at the surface and superconductivity is destroyed by weak fields, as in Onnes' 1913 experiment, and a new type of superconductor, now called a type II superconductor, capable of carrying bulk supercurrent at high fields. The key understanding that the behaviour of type II superconductors is due to quantized magnetic vortices was achieved by Abrikosov in the 1950s.

In spite of these theoretical insights, it was not until 1961 that Kunzler *et al.*² showed that high-field applications really were possible. Drawing Sn inside a Nb wire and reacting it to form the brittle intermetallic Nb_3Sn , they reached a current density $J > 10^5 \text{ A cm}^{-2}$ at 8.8 T and 4.2 K. This astounding result demonstrated that superconductivity could indeed exist in very high magnetic fields. In fact, Onnes's vision for a 10 T, high-field magnet has been abundantly fulfilled. Laboratory superconducting magnets exist by their thousands, some producing fields exceeding 20 T. Fermilab, Brookhaven, DESY and CERN all have accelerators

composed of kilometres of superconducting bending and focusing magnets (see, for example, the proceedings of the biennial Applied Superconductivity Conferences published in odd-year volumes of *IEEE Trans. Magn.* up to 1989, and *IEEE Trans. Appl. Supercond.* since 1991). The medical technique of magnetic resonance imaging was developed using very homogeneous, persistent-mode superconducting magnets, a business now exceeding US\$3 billion per year.

All of these applications are based on two low-temperature superconducting (LTS) materials, Nb-Ti alloy ($T_c = 9 \text{ K}$) or Nb_3Sn ($T_c = 18 \text{ K}$). For such applications, superconductivity is a true enabling technology and the use of liquid helium or helium-driven refrigerators is immaterial. But helium cooling tends to be too expensive for replacement of conventional electrotechnology based on Cu and Fe, two cheap and well understood materials. The first period of superconducting applications, from approximately 1962 to 1986, explored various power applications but found them too complex and expensive, thus relegating superconductivity to new high-technology and medical applications where no competitors existed. By 1986, the highest- T_c compounds possessed the A15 structure, the lightest of which, Nb_3Ge , has a T_c of 23 K (ref. 3), 5-K higher than for Nb_3Sn .

Prospects changed dramatically in late 1986, when Bednorz and Muller⁴ discovered superconductivity at 30 K in the layered cuprate, $\text{LaBa}_2\text{CuO}_{4-x}$. In early 1987, superconductivity in $\text{YBa}_2\text{Cu}_3\text{O}_x$ (YBCO) at 92 K was announced⁵, well above the boiling point of liquid nitrogen (77 K). Soon T_c rose to more than 130 K (ref. 6) in $\text{Hg}_x\text{Ba}_2\text{Ca}_2\text{Cu}_3\text{O}_x$. Widespread replacement of Cu and Fe by these new high-temperature superconductors (HTSs) was broadly discussed and significant public and private programmes to build electric utility superconducting devices were put in place in Japan, Europe and the United States⁷. Major components of the generation, transmission (power cables and devices for superconducting magnetic energy storage), distribution (transformers and fault current limiters) and end-use (motor) devices have been built, primarily using the $(\text{Bi,Pb})_2\text{Sr}_2\text{Ca}_2\text{Cu}_3\text{O}_x$ (Bi-2223) conductor⁷.

Superconductivity can have a significant role in deregulated electricity markets and in lessening CO_2 emissions and

other environmental impacts, but significant market penetration of HTS devices requires HTS wires that fully exploit their fundamental current-carrying capability. The announcement by Akimitsu in January 2001 that the binary compound MgB_2 superconducts up to 39 K (ref. 8) has generated new interest in superconductors for power applications, owing to the cheap abundance of Mg and B and the potential of analogous compounds.

Conductor requirements for power technology

A conductor is more than just the superconductor. Conductors for power applications are multifilamentary wires or tapes in which many superconducting filaments are embedded in a matrix of a normal metal, such as Cu or Ag, which provides protection against magnetic flux jumps and thermal quenching⁹ (Fig. 1). Such wires must have sufficient strength to withstand the fabrication process, device winding, cool-down and electromagnetic stresses, and be capable of being made or cabled to sufficient size to carry operating currents from hundreds to thousands of amperes at costs comparable to Cu. Estimates of the acceptable cost range from about US\$10 to US\$100 per kA m, the scale being set at its lower end by the cost for Cu (~US\$15 per kA m) and at the upper end by higher-cost applications where superconductivity has advantages not possessed by present technology^{7,10}. Such higher-cost applications include high power density underground power cables in inner cities, environmentally friendly, oil-free HTS transformers, or superconducting magnetic energy-storage devices for power networks where low reactance and instantaneous redistribution of power is vital¹¹.

These requirements define a parameter set that restricts present choice to Bi-2223 or YBCO. The combination of critical current density J_c , field and operating temperature is summarized in Table 1. Target superconductor current densities must attain 10^4 – 10^5 A cm⁻² in fields of 0.1–10 T at temperatures of 20–77 K. The data in Table 1 represent an ongoing dialogue between the R&D and applied communities that tends to push up the operating temperature, current density and field so as to better separate a Cu and Fe technology limited to fields of 1–2 T from a higher-field, superconductor-based technology. Reliable and inexpensive refrigeration is a critical part of the HTS technology and significant advances have been made in this area too¹².

At present, the only HTS conductor in production is Ag-sheathed Bi-2223 with $T_c \sim 108$ K (its lower- T_c sibling $\text{Bi}_2\text{Sr}_2\text{CaCu}_2\text{O}_x$ has only a small application in the power sector). Figure 1 shows a cross-section of a Bi-2223 conductor of the type being used for the Detroit Edison 100-MVA power cable, in which multiple strands are wound to make a fully transposed 6,000-A cable operating at 70–75 K. The individual strand is 0.2 mm by 4 mm with a critical current I_c of the order of 125 A in zero field at 77 K. Underpinning conductor R&D programmes worldwide is the conviction that such conductors have neither the magnetic field performance for applications at liquid nitrogen temperatures (~65–80 K), nor sufficiently low costs. The potential lower-cost alternative is a biaxially textured YBCO conductor or possibly MgB_2 or an analogous compound¹³. We now explore the underlying materials issues that define these beliefs.

Superconducting properties, crystal structure and anisotropy

Figure 2 presents the important magnetic field–temperature (H – T) phase diagram for the three actual (Nb–Ti, Nb_3Sn and Bi-2223) and

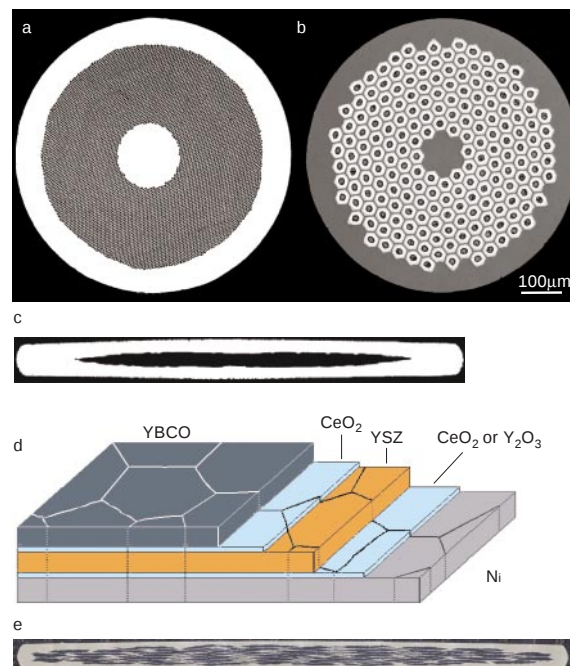


Figure 1 Conductor forms of practical superconductors. **a**, Conductor containing about 3,000 Nb47wt%Ti filaments embedded in a copper stabilizer, which protects the conductors during a transition (quench) to the normal state. The wire is ~0.8 mm in diameter and the filament diameter is 10 μm. **b**, Powder-in-tube Nb_3Sn conductor made by extruding and drawing a mixture of NbSn₂ and Cu inside the Nb tubes that form each of the 192 filaments. The conductor is ~1 mm in diameter and the filaments are ~20 μm in diameter. **c**, Cu-sheathed MgB_2 tape made by rolling MgB_2 powder in a copper tube. The tape is ~0.2 mm thick by 4 mm wide⁸⁰. **d**, Structure of a deformation-textured coated conductor. Typical thicknesses of the Ni, $\text{CeO}_2/\text{Y}_2\text{O}_3$, YSZ, CeO_2 and $\text{YBa}_2\text{Cu}_3\text{O}_x$ (YBCO) layers are (respectively) 50–100 μm, 10 nm, 300 nm, 10 nm and 0.3–2 μm. **e**, Cross-section of a 55-filament $(\text{Bi,Pb})_2\text{Sr}_2\text{Ca}_2\text{Cu}_3\text{O}_x$ (Bi-2223)-power cable tape encased in a silver matrix, which acts as stabilizer. The tape is ~0.2 × 4 mm wide. All conductors except the coated conductor were made by conventional extrusion, wire drawing and/or rolling techniques.

two potential (YBCO and MgB_2) conductor materials. Their different phase diagrams result from their distinctly different physical parameters and crystal structures, as shown in Fig. 3 and Table 2. All five are type II superconductors for which bulk superconductivity exists up to an upper critical field $H_{c2}(T)$, which can exceed 100 T for Bi-2223 and YBCO. In fact, applications are limited by a lower characteristic field, the irreversibility field $H^*(T)$ at which J_c vanishes. For the isotropic cubic superconductors, Nb–Ti and Nb_3Sn , $H^*(T)$ is about $0.85H_{c2}(T)$ (refs 14,15). All three higher- T_c compounds have anisotropic layered structures, which result in significant anisotropy of the upper critical field, $\eta = H_{c2}^{\parallel}(T)/H_{c2}^{\perp}(T)$, parallel and perpendicular to the superconducting layers. Here η has values of 1 for Nb–Ti and Nb_3Sn , 2–3 for MgB_2 (refs 16,17), 5–7 for YBCO, and

Table 1 Industry consensus wire performance requirements for various utility device applications

Application	J_c (A cm ⁻²)	Field (T)	Temperature (K)	I_c (A)	Wire length (m)	Strain (%)	Bend radius (m)	Cost (US\$ per kA m)
Fault current limiter	10^4 – 10^5	0.1–3	20–77	10^3 – 10^4	1,000	0.2	0.1	10–100
Large motor	10^5	4–5	20–77	500	1,000	0.2–0.3	0.05	10
Generator	10^5	4–5	20–50	>1,000	1,000	0.2	0.1	10
SMES*	10^5	5–10	20–77	~10 ⁴	1,000	0.2	1	10
Transmission cable	10^4 – 10^5	<0.2	65–77	100 per strand	100	0.4	2 (cable)	10–100
Transformer	10^5	0.1–0.5	65–77	10^2 – 10^3	1,000	0.2	1	<10

*SMES, superconducting magnetic-energy storage.

Data supplied by R. Blaugher.

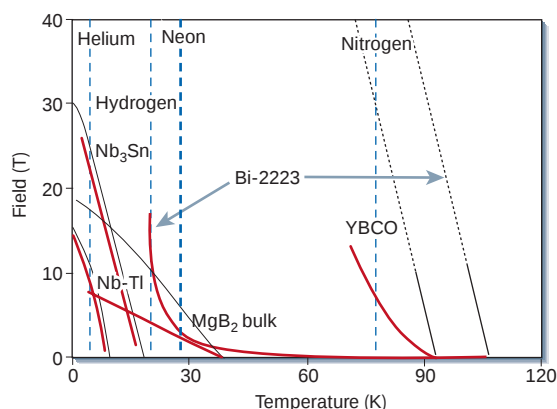


Figure 2 Magnetic field–temperature diagram for Nb47wt%Ti, Nb₃Sn, MgB₂, Bi-2223 and YBCO. The upper critical field H_{c2} at which bulk superconductivity is destroyed is indicated in black, while the irreversibility field H^* at which the bulk critical current density goes to zero is indicated in red. It can be seen that $H^*(T)$ is close to $H_{c2}(T)$ for Nb-Ti and Nb₃Sn, about half of $H_{c2}(T)$ for MgB₂, and much lower than $H^*(T)$ for YBCO and Bi-2223. There is not much scope for using Bi-2223 at 77 K because its irreversibility field (~ 0.3 T) is so low. As a result, applications at 77 K are restricted mostly to power cables for which the self-field is well below $H^*(77\text{K})$.

50–200 for Bi-2223 (ref. 18). In all cases, H_{c2} and H^* are lower for field applied parallel to the c -axis and perpendicular to the a - b planes. Figure 2 shows that strongly anisotropic Bi-2223 exhibits an enormous suppression of $H^*(77\text{K})$ to the very low value of ~ 0.2 T, well below $H_{c2}(77\text{K})$, which is of order 50 T for H parallel to the c -axis¹⁹. This low irreversibility field prevents use of Bi-2223 at 77 K in any significant field (although power cables are possible) and provides one of the key arguments for developing a second-generation HTS technology based on YBCO, for which $H^*(77\text{K}) \sim 7$ T.

The structural origin of these important superconducting anisotropies is illustrated in Fig. 3 and Table 2. Nb-Ti, normally Nb47-50wt%Ti, is a body-centred cubic, solid-solution alloy with short electron mean-free-path, high resistivity and coherence length $\xi(4\text{K}) = 5$ nm. Nb₃Sn has $\xi(4\text{K}) = 3$ nm and possesses the cubic A15 crystal structure, in which three orthogonal chains of Nb atoms separate the Sn atoms lying at cube corners and cube centre. MgB₂ has the hexagonal space group $P6/mmm$ composed of alternating B and Mg sheets, the B being nested in the Mg interstices⁸. It can form as a low-resistivity ($\rho(40\text{K}) = 0.4 \mu\Omega \text{ cm}$), perhaps perfectly stoichiometric compound²⁰ for which $\xi(4\text{K})^{\perp} = 6.5$ nm and $\xi(4\text{K})^{\parallel} = 2.5$ nm, where the perpendicular direction along the c -axis is taken with respect to the B planes in MgB₂ and the CuO₂ planes in HTSs. YBCO is a complex, layered perovskite centred on a Y layer, around which are stacked the CuO₂ plane of strong superconductivity and a double charge-reservoir layer of O-Ba-O and O-Cu. In Bi-2223, the charge-reservoir layer consists of a double (Bi,Pb)-O and its neighbouring Sr-O layers, the superconducting layers being a stack of three CuO₂ layers interleaved by two Ca layers. The anisotropy of Bi-2223 is so

large that $H^*(77\text{K})$ is only 0.2–0.3 T, restricting 77-K applications to low self-field uses such as power cables, even though $H_{c2}(77\text{K})$ exceeds 10 T.

Flux pinning and the critical current density

Under equilibrium conditions, magnetic flux penetrates the bulk of a type II superconductor above the lower critical field H_{c1} , which is of order 10–20 mT for the materials under consideration. Over most of the available H - T space, $H > H_{c1}$, this magnetic flux exists as a lattice of quantized line vortices or fluxons^{18,21}. Each fluxon is a tube of radius of the London penetration depth $\lambda(T)$, in which superconducting screening currents circulate around a small non-superconducting core of radius $\xi(T)$, where $\xi(T)$ is the superconducting coherence length. The flux carried by the screening currents in each fluxon equals the flux quantum $\phi_0 = 2 \times 10^{-15}$ Wb. Bulk superconductivity is destroyed when the normal cores overlap at the upper critical field, $H_{c2}(T) = \phi_0 / 2\pi\mu_0\xi(T)^2$. In isotropic materials such as Nb-Ti and Nb₃Sn, vortex lines are continuous, but the weak superconductivity of the blocking layers of HTS compounds produces a stack of weakly coupled ‘pancake’ vortices whose circulating screening currents are mostly confined within the superconducting CuO₂ planes¹⁸.

Superconductors can carry bulk current density only if there is a macroscopic fluxon density gradient²², defined by the Maxwell equation $\nabla \times \mathbf{B} = \mu_0 \mathbf{J}$. This gradient can be sustained only by pinning the vortices (flux pinning) at microstructural defects. Increasing T and H weaken the potential wells at which vortices are pinned, thus lessening $H^*(T)$ and $J_c(T, H)$. Flux pinning is determined by spatial perturbations of the free energy of the vortex lines due to local interactions of their normal cores and screening currents with these microstructural imperfections²². In addition, the fluxon structure is subjected to the Lorentz force $\mathbf{F}_L = \mathbf{J} \times \mathbf{B}$ of the macroscopic current. The critical current density $J_c(T, H)$ is then defined by the balance of the pinning and Lorentz forces, $F_p = F_L$, where F_p is the volume summation over all microstructural pinning defects in the strongly interacting network of flux lines. Ideally, a type II superconductor can carry any non-dissipative current density J smaller than J_c . When J exceeds J_c , a superconductor switches into a dissipative, vortex-flow state driven by the Lorentz force.

This description of flux pinning immediately suggests tailoring the defect structure of the conductor to maximize J_c . In fact, this does not correspond well to the essentially empirical way in which conductors have been developed, especially those of Bi-2223 and YBCO, for which the important flux-pinning defects are largely unknown. In practice, it is the critical current I_c of a wire of cross-section A that is measured, normally at the finite electric field of $1 \mu\text{V cm}^{-1}$. The physics of vortex dynamics and pinning is enormously interesting^{18,22}, but equally important in practical terms is the fact that the actual cross-section over which transport currents flow in HTS conductors is much less than the whole²³. What has greatly held back the development of Bi-2223 conductors in particular is uncertainty about whether the long-range utilization of the current-carrying cross-section of state-of-the-art conductors is 5, 25 or 50%. It is certain that it is not 100%, for reasons discussed later. Irrespective of such percolative current-flow effects, the upper limit to the critical current is set by

Table 2 Basic material and critical current density relevant parameters for practical superconductors

Material	Crystal structure	Anisotropy	T_c (K)	H_{c2}	H^*	In-plane coherence length $\xi(0)$ (nm)	In-plane penetration depth $\lambda(0)$ (nm)	Depairing current density (A cm ⁻²), 4.2 K	Critical current density (A cm ⁻²)	$\rho(T_c)$ ($\mu\Omega \text{ cm}$)
Nb47wt%Ti	Body-centred cubic	Negligible	9	12 T (4 K)	10.5 T (4 K)	4	240	3.6×10^7	4×10^5 (5 T)	60
Nb ₃ Sn	A15 cubic	Negligible	18	27 T (4 K)	24 T (4 K)	3	65	7.7×10^8	$\sim 10^6$	5
MgB ₂	$P6/mmm$ hexagonal	2–2.7	39	15 T (4 K)	8 T (4 K)	6.5	140	7.7×10^7	$\sim 10^6$	0.4
YBCO	Orthorhombic layered perovskite	7	92	>100 T (4 K)	5–7 T (77 K)	1.5	150	3×10^8	$\sim 10^7$	~40–60
Bi-2223	Tetragonal layered perovskite	~50–100	108	>100 T (4 K)	~0.2 T (77 K)	1.5	150	3×10^8	$\sim 10^6$	~150–800

a

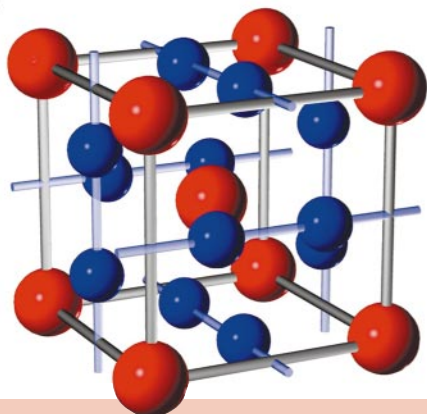
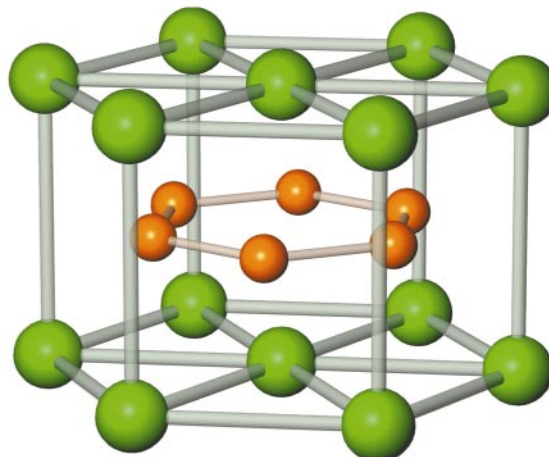
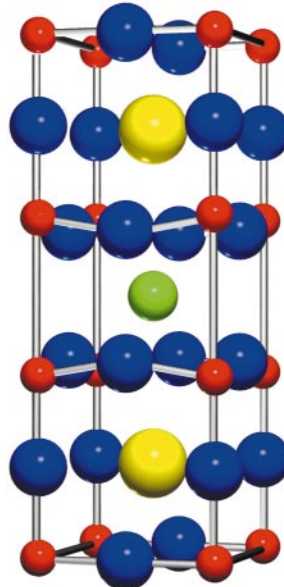
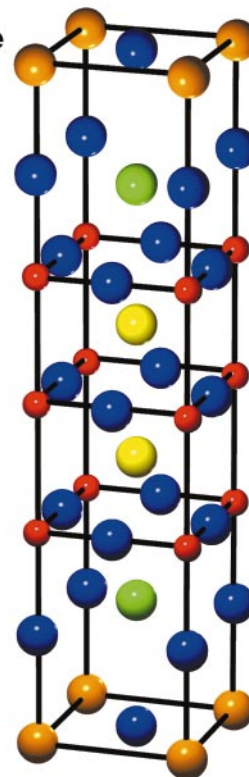
b

c

d

e


Figure 3 Crystal structures of Nb47wt%Ti, Nb₃Sn, MgB₂, Bi-2223 and YBCO. **a**, Nb47wt%Ti has the disordered body-centred-cubic structure. **b**, Nb₃Sn has the A15 crystal structure in which Sn atoms lie at cube-corner and cube-centre sites, and the Nb atoms lie in three orthogonal chains. **c**, MgB₂ has the hexagonal AlB₂ *P6/mmm* sheet structure in which B atoms lie in a completely enclosed sheet between Mg layers. **d**, YBCO has a layered, orthorhombic perovskite structure composed of two charge-reservoir layers (Cu-O, Ba-O₂) sandwiching two CuO₂ planes of strong superconductivity. **e**, Bi-2223 has a tetragonal, layered, orthorhombic perovskite structure composed of two charge-reservoir layers (Bi(Pb)-O, Sr-O) sandwiching three CuO₂ planes of strong superconductivity.

the flux-pinning current density, which cannot exceed the depairing current density $J_d = \phi_0/[3/(\sqrt{3})\pi\mu_0\lambda^2(T)\xi(T)]$, the maximum supercurrent density circulating near the vortex cores.

The best information that we possess about the limits to flux pinning in practical superconductors has come from extensive studies of Nb-Ti²⁴. Exceptionally strong pinning by a dense, ~20–25-vol% lamellar structure of normal (that is, non-superconducting) α -Ti ribbons, about 1 nm (0.2 ξ) thick and aligned parallel to the wire transport current, can produce a J_c that approaches 5–10% of J_d at zero field and 4.2 K. The flux-pinning mechanism is also known for Nb₃Sn, in which J_c is determined by the magnetic interaction of the fluxon currents with grain boundaries^{25,26}. In this case, $J_c(T,H)$ increases with decreasing grain size, thus encouraging low-temperature methods of phase formation that produce a fine grain size²⁷. There are also indications of grain-boundary pinning in MgB₂ (refs 28–30). All three materials possess zero-field J_c values, which can exceed 1 MA cm⁻² at 4.2 K.

The important flux-pinning interactions in conductor forms of YBCO and Bi-2223 are largely unknown, partly because the coherence length is so small that even atomic-sized point defects can pin fluxons. The critical temperature of HTSs is extremely sensitive to the carrier

(hole) density, which is in turn determined by local oxygen (and other) non-stoichiometries, so that even a weak hole-depletion at crystalline defects can locally

drive an HTS into an antiferromagnetic insulator. The proximity of the superconducting state to the metal-insulator transition, the d-wave pairing symmetry, the very short in-plane coherence length ($\xi(0) = 1.5$ –2 nm) and the much longer Debye screening length ($l_D \sim \xi(0)$) all combine to produce significant suppression of the superconducting order parameter $\Delta(\mathbf{r})$ near crystal defects such as impurities, dislocations and grain boundaries³¹, thus making them effective pinning centres. In fact, recent scanning tunnelling microscopy of Bi-2212 single crystals³² has revealed significant variations of $\Delta(\mathbf{r})$ near single-atom impurities (Zn and Ni), as well as striking intrinsic variations of $\Delta(\mathbf{r})$ on scales of 1–2 nm even in pure crystals³³.

This extreme sensitivity of HTSs to nanoscale defects, along with the intrinsic spatial fluctuations of $\Delta(\mathbf{r})$ from 25 meV to 60 meV on the scale of ξ or less³³, provide effective core pinning of vortices. There is strong experimental evidence that point defects such as Zn and Ni impurities³² or oxygen vacancies³⁴, or line defects such as edge³⁵ and

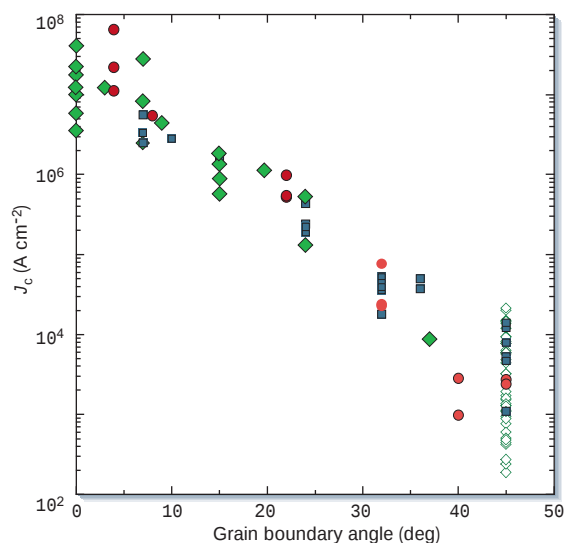


Figure 4 Transport critical current density at 4.2 K measured in thin films of YBCO grown on [001] tilt bicrystal substrates of SrTiO₃ (squares³⁷ and filled diamonds⁴⁴), Y₂O₃-stabilized ZrO₂ (circles⁹⁷), and bi-epitaxial junctions (open diamonds⁹⁸) of varying misorientation angle θ . Data of refs 97 and 44 were taken at 77 K and have been multiplied by a factor of 10.9 to make them comparable with the data at 4.2 K. Despite a significant scattering, $J_c(\theta)$ exhibits a universal exponential dependence on θ . Data compilation courtesy of J. Mannhart³⁷.

screw³⁶ dislocations, can effectively pin pancake vortices in layered HTSs. Thus, all forms of HTSs possess crystalline defects of different kinds, which provide strong flux pinning at low temperatures. Consistent with this, YBCO and Bi-2223 films^{37,38} can exhibit $J_c(4K)$ values of more than 10 MA cm⁻² and single crystals of YBCO (ref. 39) can exhibit values up to 1 MA cm⁻². Although these J_c values are about an order of magnitude lower at 77 K, they are more than sufficient for power applications (Table 1). Moreover, they are still much lower than the depairing limit: $J_d \approx 300$ MA cm⁻² at 4 K for characteristic values $\lambda(0) \approx 150$ nm and $\xi(0) \approx 1.5$ nm. Thus, unlike Nb-Ti where the material is rather close to its performance limit, there is enormous untapped potential for improved HTS material performance.

The main obstacle for HTS conductors is that their current-carrying capability deteriorates rapidly with increasing temperature and magnetic field. As shown in Fig. 2, J_c vanishes at the irreversibility field $H^*(T)$, which is far below $H_{c2}(T)$ at 77 K. An independent, significant limit on J_c is current blockage by grain boundaries, as described in the next section. But the underlying cause of the suppression of $H^*(T)$ at higher temperatures is due to the layered structure of HTSs, which greatly facilitates depinning of stacks of weakly coupled pancake vortices by thermal fluctuations^{15,18}. For $H < H^*(T)$, vortex fluctuations cause thermally activated creep of magnetic flux⁴⁰, producing measurable dissipation well below J_c defined at $1 \mu V cm^{-1}$. Anisotropy strongly enhances the effect of thermal vortex fluctuations¹⁸, thus suppressing H^* and enhancing flux creep in Bi-2223 much more than in YBCO. Both H^* and J_c can be significantly improved by irradiation, which creates effective columnar pinning tracks⁴¹. For example, neutron irradiation of U-doped Bi-2223 tapes fissions the U and produces heavy-ion tracks that raise $H^*(77K)$ to over 1 T, while also reducing the anisotropy of J_c (ref. 42).

Grain boundaries

Large devices need kilometres of polycrystalline conductors and the influence of grain boundaries is therefore critical. The most unambiguous evidence that grain boundaries are strong barriers to current

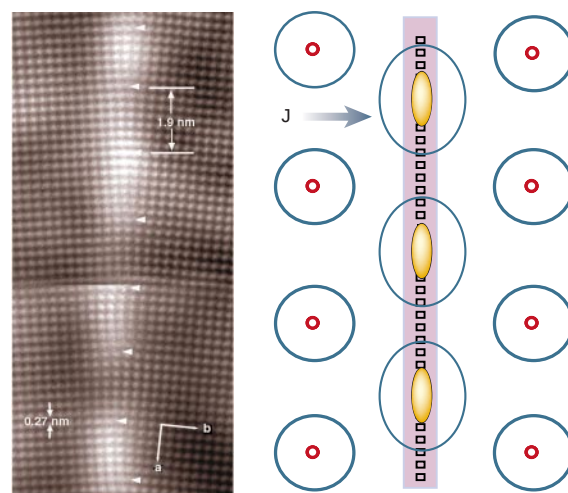


Figure 5 Grain boundary structure and its effect on vortex properties. The left panel is a Fourier-filtered transmission electron micrograph of an 8° [001] tilt Bi₂Sr₂CaCu₂O_x bicrystal, showing the edge dislocation structure at the grain boundary and the channels of good conduction that lie between their cores. Stacking faults are indicated by the white regions between alternate pairs of dislocations. Strong strains are found at the core of each dislocation. On the right is a schematic illustration of the consequences of the dislocation structure for vortices at the boundary. Abrikosov vortices are indicated within the grains, but their cores become elongated when they sit at the boundary, owing to the reduced current transparency of the boundary. The grain-boundary dislocations are indicated by the black squares and the carrier-depleted zone by the pink-shaded region. Vortices at the grain boundary acquire a mixed Abrikosov-Josephson character, which strongly reduces their pinning force along the boundary.

flow in HTSs comes from detailed studies of [001] tilt YBCO bicrystals on SrTiO₃ or Y₂O₃-stabilized ZrO₂ (YSZ) substrates. These have shown that the critical current density J_b across the grain boundary drops exponentially below that of the grains, $J_b = J_0 \exp(-\theta/\theta_c)$, as a function of the misorientation angle θ between the neighbouring crystallites^{37,43–45}, where $\theta_c \approx 2–5^\circ$, depending on the value of the intragrain J_c (Fig. 4). This extreme sensitivity to misorientation, coupled with the intrinsic anisotropy of the HTS compounds, dictates the need to texture conductors into the tape forms shown in Fig. 1, so as to shift the grain-boundary misorientation distribution to as small a value of θ as possible. For Bi-2223, the rolling deformation used to make the tape produces a marked uniaxial, *c*-axis texture, whereas for YBCO-coated conductors, the essential goal is to make both in- and out-of-plane texture so that the tape behaves as a quasi-single crystal, even though the YBCO grain size is $< 1 \mu m$ in diameter^{46,47}.

This behaviour of HTS materials is in strong contrast to metallic LTS and MgB₂, in which grain boundaries are not only transparent to current, but also significantly contribute to flux pinning, increasing the overall J_c as the grain size decreases^{26,28,48}. Although Table 2 might suggest that HTS grain boundaries are weak links because of their short coherence length, in fact a value of $\xi(77K)$ of ~ 4 nm for YBCO is actually longer than for Nb₃Sn at 4 K. Moreover, low carrier density makes grain boundaries in (Ba,Pb)BiO₃ also weak-linked, even though $\xi(4K) \sim 7$ nm for this compound^{49,50}. In fact, the weak-link (that is, current-obstructing) behaviour of grain boundaries in HTSs results from the same factors, which otherwise enhance the pinning of vortices by point defects, namely the strong dependence of T_c on hole concentration and low carrier density. A tilt grain boundary is built of a chain of edge dislocations spaced by $d = b/2\sin(\theta/2)$, where

$b \approx 0.4$ nm is the Burgers vector. As the misorientation angle θ increases, the spacing between the insulating dislocation cores decreases, becoming of the order of the coherence length ξ at $\theta \approx 5-7^\circ$. For higher angles, a grain boundary becomes a Josephson weak link, because of the local suppression of the order parameter in the current channels between the dislocations, and J_b decreases exponentially with increasing θ ³¹. Extensive electron microscopy has established that the hole content is depressed in the vicinity of grain boundaries and that there is only partial occupancy of cations in the structural units that form the boundary^{51,52}.

Grain boundaries possess extra ionic charge, whose magnitude increases with increasing θ and causes hole depletion in a layer of thickness given by the Thomas-Fermi screening length $l_D \approx \xi(0)$ (Fig. 5). As a result, the grain boundary is in a locally underdoped state, the local order parameter suppression increasing as θ increases^{31,52}. That is why local overdoping by partially substituting Ca for Y in YBCO (refs 53,54) is effective. Adding extra carriers to the grain boundary with Ca-doping and reducing local strains can significantly increase J_b (by a factor of eight for a 24° [001] tilt boundary at 4.2 K and 0 T (ref. 53), and by a factor of three for a 7° -bicrystal at 55 K and 3 T (ref. 55)). Although overdoping ameliorates, rather than removes, the problem of current obstruction by grain boundaries in HTSs, it may certainly be useful for conductors^{43-45,56,57}, because the critical angle θ_c is only $2-5^\circ$.

Performance of HTS conductors in a magnetic field is determined by the nature of the vortices at the grain boundaries and their pinning interaction with structural defects, and by their magnetic interactions with strongly pinned vortices within the grains (Fig. 5). Because of the extended core of grain-boundary vortices, they are generally pinned more weakly than vortices in the grains (ref. 58, and Gurevich *et al.*, submitted), and grain boundaries become barriers to current flow because their critical current density $J_b(T, H)$ is smaller than $J_c(T, H)$ in the grains. This behaviour is clearly visible in the magneto-optical images of Fig. 6, which shows preferential flux penetration along a network of coated-conductor grain boundaries⁵⁶ having $\theta > 4^\circ$.

The influence of misorientation is less certain for compounds other than YBCO, as all other HTS compounds are harder to grow as good films on bicrystal substrates. To first approximation, thin films of $\text{Bi}_2\text{Sr}_2\text{CaCu}_2\text{O}_x$ and Bi-2223 show similar exponential decrease of $J_b(\theta)$ (refs 59,60) to that seen for YBCO, suggesting a common mechanism for suppression of critical current on grain boundaries. Experiments on float-zone $\text{Bi}_2\text{Sr}_2\text{CaCu}_2\text{O}_x$ bicrystals indicate much weaker dependence of J_b on θ than for thin films⁶¹, but the intragranular critical current densities are so small that they might not reach the critical current density of the grain boundaries. An extensive magneto-optical study of [001] tilt, thin-film YBCO bicrystals⁶² has shown that the temperature dependences of J_c and J_b are significantly different.

Current percolation in polycrystals

Any conductor must be polycrystalline. All conductor materials except Nb-Ti are brittle, making cracks endemic in Nb_3Sn , MgB_2 , Bi-2223 and YBCO. In addition, the last three compounds all are prone to porosity. The net result is that current percolates through a polycrystalline network containing many obstructions, some of which partially block the current while others result in a total block²³. Thus the local fraction of current-carrying cross-section is significantly less than unity. Magneto-optical imaging is an effective method of visualizing non-uniformities of current flow in polycrystalline conductor forms, as shown in Fig. 7 for MgB_2 , Bi-2223 and YBCO. These images show the 'roof' pattern of the perpendicular magnetic field, resulting from screening currents induced by an external field. For MgB_2 the finest, visible-scale variation seems to be controlled by $\sim 10\%$ porosity⁶³ and not by grain boundaries^{17,20,64}. For the Bi-2223 tape, a transverse feathering of the image is caused principally by quasiperiodic fluctuations of the superconductor

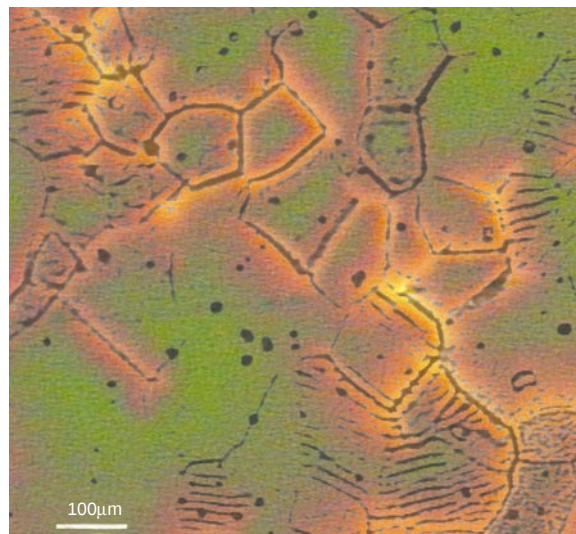


Figure 6 Magneto-optical image of the flux penetrating into a typical deformation-textured YBCO coated conductor overlaid on a light-microscope image of the underlying Ni substrate. Darker (green) areas are regions that are well connected electromagnetically, whereas the lighter (orange) flux network indicates where the local current density is reduced. Some grain boundaries in the Ni substrate (black) do not appear in the magneto-optical image. However, all grain boundaries with misorientations greater than 4° do appear⁵⁶.

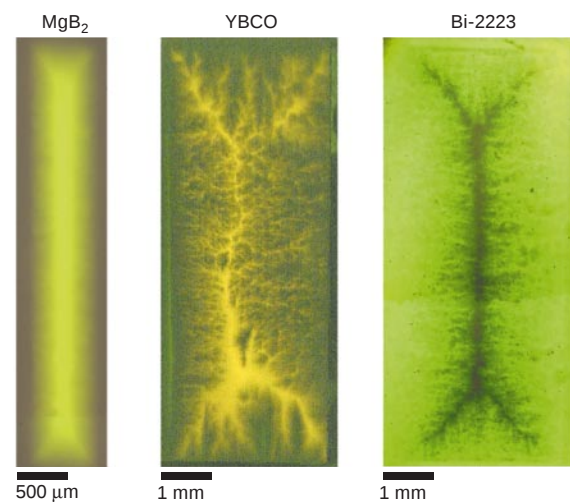
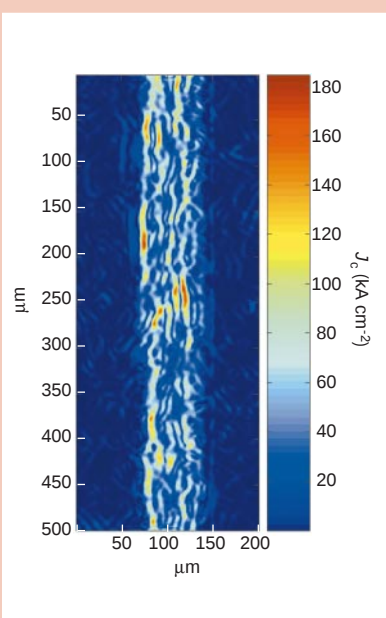


Figure 7 Magneto-optical images of (left to right) a sintered MgB_2 slab⁶³, a YBCO IBAD-coated conductor, and a Bi-2223 monocore tape²³. Each tape shows a global roof-top pattern of the perpendicular component of magnetic field above the superconductor surface. This 'roof-top' pattern arises from induced magnetization currents circulating through the entire sample. The MgB_2 slab shows the most uniform current flow, even though it is imaged in an applied field of zero after cooling in 120 mT to 37 K, only 2 K below T_c . Almost no spatial variation of J_c is visible. The coated conductor was imaged after field-cooling in 60 mT to 11 K, well below its T_c of 90 K, then reducing the applied field to zero. The Bi-2223 tape was imaged at 11 K after cooling without field and then applying a field of 120 mT. Some feathering of the image transverse to the tape length indicates a longitudinal variation of the current, probably associated with rolling defects.

Figure 8 Colour map of the spatial distribution of the local critical current density in the same monocore Bi-2223 tape imaged in Fig. 7. The tape was imaged in a slab geometry produced by sectioning the tape along a centre line parallel to the tape length. The image from which the current reconstruction was performed was made by cooling the slab to 77 K in zero field, then applying a field of 36 mT parallel to the half width of the slab. The transport critical current density $J_c(77\text{K}, 0\text{T})$ of the tape was 35 kA cm^{-2} , about the median of the distribution seen in the J_c map. Peak values of the J_c exceeded 180 kA cm^{-2} .



thickness and residual rolling damage²³. For the YBCO-coated conductor, there are multiple lines emanating from the central roof pattern, indicating local, higher current density loops superimposed upon the long-range current. Thus magneto-optical examination shows explicitly that there are multiple scales of current loops flowing in polycrystalline HTS samples. Such images pose the question: by how much is the average current density — defined by $J_c = I_c/A$, where A is the total cross-section, rather than the active cross-section $A_{\text{effective}}$ — reduced below the flux-pinning critical current density established within each grain?

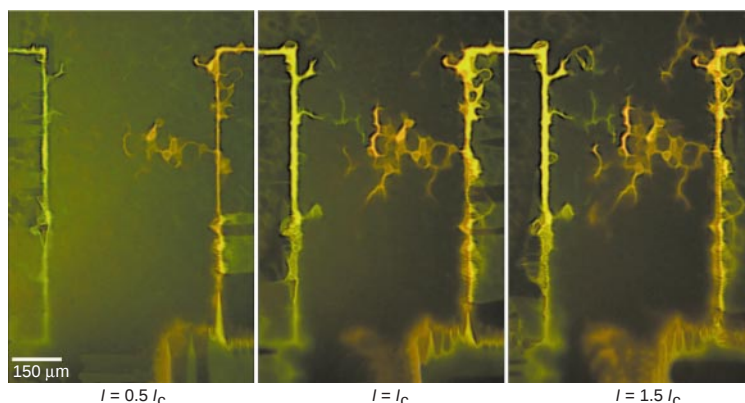
To answer such a question requires good understanding of the intragrain critical current densities. In fact, the magnitude of the intragrain J_c of Bi-2223 remains unclear, because few high-quality bulk or thin-film single crystals have been made. Several thin-film results^{38,60} suggest that $J_c(77\text{K}, 0\text{T})$ can achieve 1 MA cm^{-2} , whereas a single-crystal study⁶⁵ gave J_c values significantly less than 10^5 A cm^{-2} . At 77 K and 0 T, Bi-2223 conductors can achieve up to $\sim 75\text{ kA cm}^{-2}$ for small current conductors⁶⁶ and $40\text{--}50\text{ kA cm}^{-2}$ for high current conductors having $I_c \sim 150\text{ A}$ (refs 19, 67). Figure 8 shows the results of a recent magneto-optical current reconstruction²³ on a high- J_c monofilament conductor with $J_c(77\text{K}, 0\text{T}) = 35\text{ kA cm}^{-2}$. Remarkably, J_c can achieve local values of 180 kA cm^{-2} , up to five times higher

than the transport J_c . Although there is a tendency for the highest- J_c regions to be located at the Ag-superconductor interface, in fact very-high- J_c regions are located throughout the tape. Such local variability is not surprising for a seven-component system, because it is hard to control the phase-conversion reaction of $\text{Bi}_2\text{Sr}_2\text{CaCu}_2\text{O}_x$ and other constituents to Bi-2223. Images such as Fig. 8 offer the possibility of directly correlating the local microstructure to the local J_c . It is interesting to note that regions of J_c exceeding 100 kA cm^{-2} are $50\text{--}100\text{ }\mu\text{m}$ long, several times the $\sim 20\text{-}\mu\text{m}$ grain length. Although earlier work⁶⁸ suggested that high- J_c regions were confined mainly to the Ag interface region, more recent results show that better processing is producing much more uniform J_c distribution across Bi-2223/Ag tapes, resulting in significant gains in tape performance.

State-of-the-art Bi-2223 conductors contain grains $\sim 20\text{ }\mu\text{m}$ long by $\sim 1\text{ }\mu\text{m}$ thick, distributed in $30\text{--}60$ filaments, each of which is of the order of $5\text{--}10\text{ }\mu\text{m}$ thick by $200\text{--}300\text{ }\mu\text{m}$ wide. The current percolates in a three-dimensional manner through a polycrystalline array, which possesses a significant c -axis texture, of the order of $10\text{--}12^\circ$ full-width at half-maximum (FWHM), but has little in-plane texture. By contrast, YBCO-coated conductors rely on epitaxy and are controlled mostly by the two-dimensional percolation^{69,70}. Figure 9 images the local field produced by current flowing along a track in a representative coated conductor. An extended planar array of misoriented grains obstructs the current, forcing it to flow non-uniformly and causing local dissipation, even at $0.5I_c$, where, as usual, I_c is defined at the electric field criterion of $1\text{ }\mu\text{V cm}^{-1}$ established across the whole track. As I increases, the current-obstructing network expands, eventually becoming continuous across the whole conductor. Such images show that I_c is determined locally, with many regions of the sample showing no signs of dissipation even at $I = 1.5I_c$. A similar conductor had a global $J_c(77\text{K}, 0\text{T})$ of 1.2 MA cm^{-2} , but individual clusters of YBCO grains had J_c values⁵⁷ exceeding 5 MA cm^{-2} . Thus, even in the biaxially textured coated conductors for which both in- and out-of-plane texture FWHM values are significantly below 10° , there is abundant evidence for current percolation around a network of obstructing grain boundaries and other defects.

Present $J_c(77\text{K}, 0\text{T})$ values in $1\text{--}2\text{-cm}$ -long coated conductor prototypes^{71–74} typically fall in the range $1\text{--}3\text{ MA cm}^{-2}$. Further performance improvement may be impeded by electromagnetic and thermal instabilities, such as flux jumping, hot-spot formation and quench propagation, the same factors which limit the current-carrying capability of low- T_c superconductors^{9,75}. Whatever future microstructure improvements occur, current-obstructing obstacles are inevitable in any practical long-length forms. Because the electromagnetic and thermal stability of the current-carrying state is influenced by hot spots near obstacles, the architecture of coated

Figure 9 Magneto-optical images at 77 K of the self-field produced by an applied transport current in a RABiTS coated conductor sample for which the full-width $J_c(77\text{K}, 0\text{T})$ was 0.7 MA cm^{-2} . A laser was used to cut a link restricting the current flow to a region $0.5 \times 1.1\text{ mm}$, which had a $J_c(77\text{K}, 0\text{T})$ of 0.6 MA cm^{-2} . Images shown are at applied currents for the link of (left to right) $0.5I_c$, I_c , and $1.5I_c$, where I_c is defined at $1\text{ }\mu\text{V cm}^{-1}$. A cluster of YBCO grain boundaries is visible in all the images, indicating where in the link J_c is limited well before the onset of bulk dissipation. As the current is increased, percolation near the visible cluster of grain boundaries becomes more pronounced, but above and below this cluster current flows more uniformly, and preferentially near the edges of the link. Even at $1.5I_c$, much of the link is supporting a current less than its local J_c . Patterning of bridges has shown that the intragrain J_c can reach 5 MA cm^{-2} , several times the J_c of this bridge. It is clear that current is percolating through a network of grain boundaries, which also result in localized dissipation in the track.



conductors needs to be optimized so as to reduce local dissipation by current redistribution to the Ag stabilizer, and enhance heat transfer from the YBCO to the coolant. An important question is what is the acceptable size and concentration of obstructive hot spots, which provides stable conductor operation.

Because obstructions initiate dissipation, the spatial distribution of the electric field $E(x,y)$ near such obstructions controls the J_c of the conductor. But unlike current distributions, which can be extracted from magneto-optical images, the electric field distribution is much harder to measure. Recently, analytical methods for the calculation of $E(x,y)$, taking account of the highly nonlinear E – J characteristic of a superconductor, have been developed^{76,77}. For $J < J_c$, the E – J relation is determined by thermally activated flux creep, which is conventionally approximated as $E = E_c(J/J_c)^n$, where $E_c \sim 1 \mu\text{V cm}^{-1}$, and the exponent $n(T,H)$ varies from 20–30 for $H \ll H^*$ to $n \approx 1$ at $H = H^*$. Calculations for a planar defect perpendicular to current flow show that the nonlinearity of $E(J)$ strongly enhances the electric field in extended regions of length $L_\perp \sim na$ perpendicular to the current, on a scale much larger than the defect size a . This behaviour is in stark contrast to ohmic conductors, in which current flow becomes essentially uniform at a distance $\sim a$ away from the defect. These long-range disturbances of $E(x,y)$ become particularly important in a coated conductor, because even a small defect can produce a large hot spot that spans the entire conductor if $a > d/n$, where d is the sample width. If this rather weak condition is satisfied (typically, $n \sim 20$ –30 in HTSs), then even small defects produce significant local excess dissipation. Figure 10 shows the calculated distribution of the electric field near a planar defect of length $a = 0.1d$ and $n = 20$. Even for such a small defect, the disturbance of $E(x,y)$ is much stronger than the applied field E_0 and extends all the way to the opposite side of the film⁷⁷.

Materials fabrication considerations

One of our fundamental conclusions is that both Bi-2223 and YBCO fulfil the essential requirements for wide applications of superconductivity in the electric utility network. Conductors of Bi-2223 (and Bi₂Sr₂CaCu₂O_x) are available today from several companies for service in ‘real-world’ utility sites. Cost and performance are not yet as good as are needed for widespread substitution of copper and iron, but significant special applications exist even at much higher costs. A key question for a cost-sensitive market for HTSs is to understand how much residual J_c performance capability there is in the materials themselves. As we have emphasized, the intragrain J_c capability of both Bi-2223 and YBCO is several times higher than realized in today’s conductors. Material defects control the overall critical current density of present conductors and these defects are intimately tied to their particular fabrication processes. Although the specifics of particular fabrication processes are beyond the scope of this article, certain general points can be made to emphasize the decisive role that materials engineering now must play in developing this technology.

As Figure 1 shows, four of the five conductors are made by conventional composite metal-working techniques. A merit of conventional metal working is that there is little conceptual barrier to scale-up. Making longer wires is mostly a question of starting with larger pieces and using larger equipment. Both MgB₂ and Bi-2223 wires or tapes use essentially the same technology as that used for Nb₃Sn powder-in-tube composites. In fact, there is now no fundamental barrier to the production of a kilometre-long MgB₂ wire, and many groups have already demonstrated powder-in-tube conductors^{30,78–82}. What is less clear is how much of a niche there is for MgB₂ to fill, because even in its high resistivity form which maximizes the upper critical field¹⁷ it seems that the perpendicular irreversibility field is less than for Nb₃Sn. This is likely to restrict MgB₂ to applications of lower field, say 2–3 T, where the ability to operate at ~ 20 K is decisive economically¹³. By contrast, it has proven hard to scale up the YBCO-coated conductor to a genuinely continuous process, even

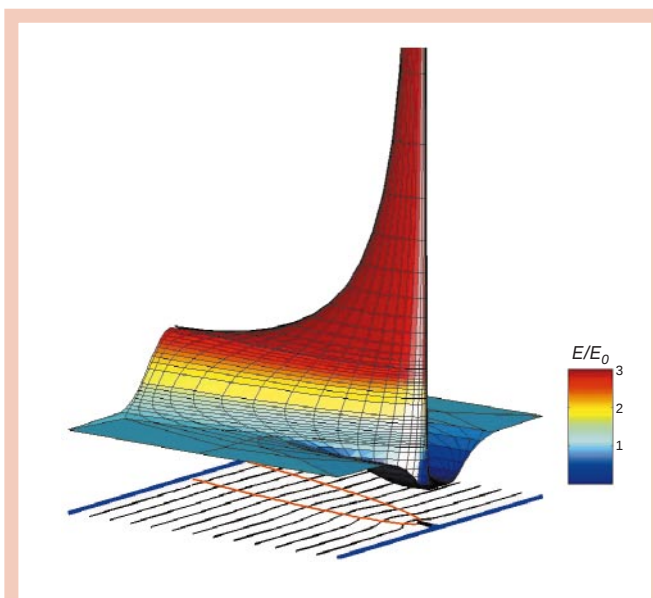


Figure 10 Spatial distribution of the electric field $E(x,y)$ for the power-law E – J characteristic, $E = E_c(J/J_c)^n$, near a planar defect of length a in a film of width d , $n = 20$ and $a = 0.1d$, as calculated in refs 76,77. The lower part of the figure shows the current streamlines (black) and the boundaries (red) of the hot spot of strongly enhanced electric field lying in the film plane. It is clear that even a small defect produces the extended hot spot that effectively blocks the entire cross-section and can result in thermal instabilities⁷⁵.

after multiple proofs of principle have shown that an appropriate substrate–buffer–YBCO architecture can be made by physical vapour deposition and/or chemical methods^{71–74,83}. The key issue here is to understand the defect populations that are characteristic of each fabrication method, and the capability and cost associated with reducing them to tolerable levels.

In assessing progress on HTS conductors, it is particularly relevant to recall that Nb₃Sn is now celebrating 40 years as a conductor and is still improving^{84,85}. Bi-2223 has had 10 years of industrial development and is also still developing strongly in attaining higher current density, reducing cost and in becoming more mechanically robust^{86–88}. The complexity of its microstructure and the interactions between fabrication and basic materials issues is slowly becoming clear⁸⁹, not least because multiple companies worldwide are selling the conductor in competition with each other. As Fig. 8 shows, the key issue is to increase the connectivity of the conductor. The prospect of two- to fivefold improvements in the performance of the present process, coupled to cost reductions triggered by significant scale-up and alternative routes such as overpressure processing⁹⁰ or melt processing⁹¹, could accelerate the early penetration of superconducting technology into the utility network. At the same time, this would provide time and rationale for solving the complex issues associated with scale-up of the YBCO coated conductors.

Any manufacturing method for YBCO conductors relies on the epitaxial deposition of YBCO onto a textured template of one or more oxide buffer layers and a normal metal substrate. This template is usually created through introduction of texture either into the buffer layers by ion beam-assisted deposition (IBAD)⁴⁷, or into the metal substrate by deformation texturing using the rolling-assisted biaxially textured substrate (RABiTS)⁴⁶ approach. The IBAD method enables the use of strong, non-magnetic Ni-superalloy substrates on which an IBAD of aligned YSZ is deposited. This process can achieve a high degree of texture with a layer of YSZ that is $\sim 0.5 \mu\text{m}$ thick. But such a process is widely considered to be too slow to be commercial. IBAD of MgO produces good texture within the first 1–2 nm and thus may be rapid enough to be commercial⁹², but the degree of texture in

present IBAD-MgO-coated conductors is not yet as good as in IBAD-YSZ variants⁹³. The inclined substrate deposition (ISD)⁹⁴ process is much quicker than IBAD, but the texture is significantly worse⁹⁵. Ion-texturing processes for buffer-layer deposition⁹⁶ are receiving renewed attention owing to their simplicity, but their success depends on producing sufficient texture so to compete with the slower and more expensive IBAD process.

All of these methods of producing a textured buffer layer involve physical vapour-deposition processes, which some believe are inherently too expensive for copper and iron replacement⁷¹. An advantage of the RABiTS approach is that all buffer-layer and YBCO-coating steps can be performed with non-vacuum processes. A strong [100] cube texture is introduced into the substrate by conventional rolling and recrystallization⁴⁶. Although the RABiTS approach has been developed mainly with pure Ni, alloyed substrate materials are being developed to increase the strength and reduce the magnetism of the Ni.

At this stage it is not at all clear which route to a coated conductor will win out. The design of a coated conductor is such that a substrate thickness of $\sim 50\text{ }\mu\text{m}$ is needed to support $1\text{--}5\text{ }\mu\text{m}$ of YBCO, giving a superconductor fraction of $\sim 5\text{--}10\%$, compared with the $25\text{--}50\%$ for conductors made by metal-working techniques. Short samples of both IBAD and RABiTS conductors can both exceed 2 MA cm^{-2} for YBCO layers up to $\sim 0.5\text{ }\mu\text{m}$. But thicker YBCO layers tend to exhibit less epitaxy, more porosity and thus smaller J_c . Degradation of the J_c of the YBCO layer as it grows is a serious and poorly understood problem. The principal cause seems to be loss of epitaxy once the layer gets thicker than $\sim 0.5\text{ }\mu\text{m}$. A proper understanding is required, as virtually all growth methods except liquid-phase epitaxy seem to share this degradation.

Summary

The fundamental crystal and electronic structure of superconductors determines their structural anisotropy and superconducting properties. Generally, the higher the value of T_c , the higher the anisotropy, and the greater the upper critical field and sensitivity to nanoscale variations of the superconducting properties. This variation is positive in the case of point and line defects, because it can lead to strong flux-pinning and high critical current density. The low carrier density, short coherence length, d-wave pairing symmetry and proximity to the metal–insulator transition of HTSs makes interfaces such as grain boundaries significant barriers to current flow, except at very small misorientations. In the higher carrier density LTS materials such as Nb–Ti, Nb₃Sn, and MgB₂, grain boundaries act as beneficial flux-pinning sites without being barriers to current flow.

Anisotropy exerts two additional, important influences on conductor design, and causes significant suppression of H^* as compared to H_{c2} . MgB₂, YBCO and Bi-2223 all exhibit higher critical current density, irreversibility field and upper critical field when the magnetic field is applied parallel to their layered structure. But in practice, either because of the inevitable misalignments that are present in any real wire or because the magnetic field cannot always be arranged to be parallel to the layers, it is the poorer perpendicular properties that generally control performance. An additional burden for YBCO and Bi-2223 is that the grain-to-grain misalignment must be held to values of 5° or less if obstruction of current across the grain boundaries is to be minimized. Thus planar conductor forms, which confer texture, are preferred.

These fundamental materials issues restrict the fabrication technologies available for conductors. So far all three industrially available conductors, Nb–Ti, Nb₃Sn and Bi-2223, use composite metal-working techniques, which are relatively easy to scale up from laboratory to factory. It is already clear that this technique will work too for MgB₂. But it does not work for YBCO, for which a multilayer epitaxial technique is required. Many techniques using a variety of chemical and physical vapour-deposition techniques can be used to make short samples that demonstrate the concept of the coated

conductor. At present the coated conductor R&D community is attempting to work out which fabrication techniques work cost-effectively in continuous processes. Broad and significant applications exist and await only the resolution of the vital materials issues that control development of the cheap, high-performance conductor-fabrication technology that underpins all superconducting applications. □

- Berlincourt, T. G. Type-II superconductivity: quest for understanding. *IEEE Trans. Magn.* **23**, 403–412 (1987).
- Kunzler, J. E., Buehler, E., Hsu, L. & Wernick, J. Superconductivity in Nb₃Sn at high current density in a magnetic field of 88 kgauss. *Phys. Rev. Lett.* **6**, 89–91 (1961).
- Gavaler, J. Superconductivity in Nb–Ge films above 22 K. *Appl. Phys. Lett.* **23**, 480–482 (1973).
- Bednorz, G. & Müller, K. A. Possible high- T_c superconductivity in the Ba–La–Cu–O system. *Z. Phys. B* **64**, 189–193 (1986).
- Wu, M. K. et al. Superconductivity at 93 K in a new mixed-phase Y–Ba–Cu–O compound system at ambient pressure. *Phys. Rev. Lett.* **58**, 908–912 (1987).
- Schilling, A., Cantoni, M., Guo, J. D. & Ott, H. R. Superconductivity above 130 K in the Hg–Ba–Ca–Cu–O system. *Nature* **363**, 56–58 (1993).
- Larbaestier, D. C. et al. Power Applications of Superconductivity in Japan and Germany (World Technology and Engineering Center, Loyola College, MD, September 1997).
- Nagamatsu, J., Nakagawa, N., Muranaka, T., Zenitani, Y. & Akimitsu, J. Superconductivity at 39 K in magnesium diboride. *Nature* **410**, 63–64 (2001).
- Wilson, M. *Superconducting Magnets* (Clarendon, Oxford, 1983).
- Dew-Hughes, D. *Physics and Materials Science of Vortex States, Flux Pinning and Dynamics* NATO Science Ser. E, Vol. 356 (eds Kossowsky, R., Bose, S., Pan, V. & Durosos, Z.) 705–730 (Kluwer Academic, Dordrecht, 1999).
- Hassenzahl, W. V. Superconductivity, an enabling technology for 21st century power systems? *IEEE Trans. Appl. Supercond.* **11**, 1447–1453 (2001).
- Radebaugh, R. in *Advances in Cryogenic Engineering* Vol. 44 (eds Haruyama, T., Mitsui, T. & Yamafuji, K.) 33–44 (Elsevier Science, 1997).
- Grant, P. Rehearsals for prime time. *Nature* **411**, 532–533 (2001).
- Suenaga, M., Ghosh, A. K., Hu, Y. & Welch, D. O. Irreversibility temperatures of Nb₃Sn and Nb–Ti. *Phys. Rev. Lett.* **66**, 1777–1780 (1991).
- Brandt, E. H. The flux line lattice in superconductors. *Rep. Prog. Phys.* **58**, 1465–1594 (1995).
- Xu, M. et al. Single crystal MgB₂ with anisotropic superconducting properties. Preprint cond-mat/0105271 at <http://xxx.lanl.gov> (2001).
- Patnaik, S. et al. Electronic anisotropy, magnetic field-temperature phase diagram and their dependence on resistivity in c-axis oriented MgB₂ thin films. *Supercond. Sci. Technol.* **14**, 315–319 (2001).
- Blatter, G., Feigelman, M. V., Geshkenbein, V. B., Larkin, A. I. & Vinokur, V. M. Vortices in high-temperature superconductors. *Rev. Mod. Phys.* **66**, 1125–1388 (1994).
- Schwartzkopf, L. A., Jiang, J., Cai, X. Y., Apodaca, D. & Larbaestier, D. C. The use of the in-field critical current density, $J_c(0.1T)$, as a better descriptor of (Bi,Pb)₂Sr₂Ca₂Cu₃O_{8-x}/Ag tape performance. *Appl. Phys. Lett.* **75**, 3168–3170 (1999).
- Canfield, P. C. et al. Superconductivity in dense MgB₂ wires. *Phys. Rev. Lett.* **86**, 2423–2426 (2001).
- Tinkham, M. *Introduction to Superconductivity*. (McGraw Hill, New York, 1975).
- Campbell, A. M. & Evetts, J. E. Flux vortices and transport current in type-II superconductors. *Adv. Phys.* **21**, 194–428 (1972).
- Polyanskii, A. A. et al. Examination of current limiting mechanisms in monorec Ag/BSCCO tape with high critical current density. *IEEE Trans. Appl. Supercond.* **11**, 3269–3270 (2001).
- Meingast, C. & Larbaestier, D. C. Quantitative description of a very high critical current density Nb–Ti superconductor during its final optimization strain. II. Flux pinning mechanisms. *J. Appl. Phys.* **66**, 5971–5983 (1989).
- Kramer, E. J. Dynamics of dislocation dipole motion in the flux line lattice of type-II superconductors. *J. Appl. Phys.* **41**, 621–629 (1970).
- Dew-Hughes, D. The role of grain boundaries in determining J_c in high-field, high current superconductors. *Phil. Mag.* **55**, 459–449 (1987).
- Cooley, L. D., Lee, P. J. & Larbaestier, D. C. Changes in flux pinning curve shape for flux-line separations comparable to grain size in Nb₃Sn wires. *Adv. Cryo. Eng.* (in the press).
- Eom, C. B. et al. High critical current density and enhanced irreversibility field in superconducting MgB₂ thin films. *Nature* **411**, 558–560 (2001).
- Song, X. et al. Anisotropic grain morphology, crystallographic texture and their implications for the electromagnetic properties of polycrystalline MgB₂ thin films, sintered pellets and filaments. *Supercond. Sci. Technol.* (in the press).
- Jin, S., Mavoori, H., Bower, C. & van Dover, R. B. High critical current in iron-clad superconducting MgB₂ wires. *Nature* **411**, 563–565 (2001).
- Gurevich, A. & Pashitskii, E. A. Current transport through low-angle grain boundaries in high-temperature superconductors. *Phys. Rev. B* **57**, 13878–13893 (1998).
- Pan, S. H. et al. Imaging the effect of individual zinc impurities on superconductivity in Bi₂Sr₂CaCu₂O_{8-x}. *Nature* **403**, 746–750 (2000).
- Pan, S. H. et al. Microscopic electronic inhomogeneity in the high- T_c superconductor Bi₂Sr₂CaCu₂O_{8-x}. *Nature* **413**, 282–285 (2001).
- Kes, P. Flux pinning and creep in high-temperature superconductors. *Physica C* **185**, 288–291 (1991).
- Diaz, A., Mechin, L., Berghuis, P. & Evetts, J. E. Evidence for vortex pinning by dislocations in YBa₂Cu₃O_{7-x} low angle grain boundaries. *Phys. Rev. Lett.* **80**, 3855–3858 (1998).
- Dam, B. et al. Origin of high critical currents in YBa₂Cu₃O_{7-x} superconducting thin films. *Nature* **399**, 439–442 (1999).
- Hilgenkamp, H. & Mannhart, J. Grain boundaries in high- T_c superconductors. *Rev. Mod. Phys.* (in the press).
- Miu, L. et al. Vortex unbinding and layer decoupling in epitaxial Bi₂Sr₂CaCu₂O_{8-x} films. *Phys. Rev. B* **52**, 4553–4558 (1995).
- Vargas, J. L. & Larbaestier, D. C. Flux pinning by ordered oxygen deficient phases in nearly stoichiometric YBa₂Cu₃O_{7-x} single crystals. *Appl. Phys. Lett.* **60**, 1741–1743 (1992).

40. Yeshurun, Y., Malozemoff, A. P. & Shaulov, A. Magnetic relaxation in high-temperature superconductors. *Rev. Mod. Phys.* **68**, 911–949 (1996).
41. Civalé, L. Vortex pinning and creep in superconductors with columnar defects. *Supercond. Sci. Technol.* **10**, A11–A28 (1997).
42. Tonies, S. et al. On the current transport limitations in Bi-based high temperature superconducting tapes. *Appl. Phys. Lett.* **78**, 3851–3853 (2001).
43. Dimos, D., Chaudhari, P. & Mannhart, J. Superconducting transport properties in $\text{YBa}_2\text{Cu}_3\text{O}_{7-x}$ bicrystals. *Phys. Rev. B* **41**, 4038–4049 (1990).
44. Heinig, N. F., Redwing, R. D., Nordman, J. E. & Larbalestier, D. C. Strong to weak coupling transition in low misorientation angle thin film $\text{YBa}_2\text{Cu}_3\text{O}_{7-x}$ bicrystals. *Phys. Rev. B* **60**, 1409–1417 (1999).
45. Verebelyi, D. et al. Low angle grain boundary transport in $\text{YBa}_2\text{Cu}_3\text{O}_{7-x}$ coated conductors. *Appl. Phys. Lett.* **76**, 1755–1757 (2000).
46. Goyal, A. et al. Texture formation and grain boundary network in rolling assisted biaxially textured substrates and in epitaxial YBCO films. *Micron* **30**, 463–478 (1999).
47. Iijima, Y., Kakimoto, K., Kimura, M., Takeda, K. & Saitoh T. Reel to reel continuous formation of Y-123 coated conductors by IBAD and PLD method. *IEEE Trans. Appl. Supercond.* **11**, 2816–2821 (2001).
48. Larbalestier, D. C. & West, A. W. New perspective on flux pinning in niobium-titanium composite superconductors. *Acta Metall.* **32**, 1871–1881 (1984).
49. Cooley, L. D., Daeumling, M., Willis, T. W. & Larbalestier, D. C. Weak link behavior in polycrystalline $\text{BaPb}_{0.7}\text{Bi}_{0.25}\text{O}_3$. *IEEE Trans. Magn.* **25**, 2314–2316 (1989).
50. Takagi, T., Chiang, Y. M. & Roshko, A. Origin of grain-boundary weak links in $\text{BaPb}_{1-x}\text{Bi}_x\text{O}_3$ superconductor. *J. Appl. Phys.* **68**, 5750–5758 (1990).
51. Babcock, S. E. & Vargas, J. L. The nature of grain boundaries in high- T_c superconductors. *Annu. Rev. Mater. Sci.* **25**, 193–222 (1995).
52. Browning, N. D. et al. The atomic origin of reduced critical currents at [001] tilt grain boundaries in $\text{YBa}_2\text{Cu}_3\text{O}_7$ thin films. *Physica C* **294**, 183–193 (1998).
53. Schmehl, A. et al. Doping induced enhancement of the critical currents of grain boundaries in $\text{YBa}_2\text{Cu}_3\text{O}_7$. *Europhys. Lett.* **47**, 110–115 (1999).
54. Hammerl, G. et al. Enhanced supercurrent density in polycrystalline $\text{YBa}_2\text{Cu}_3\text{O}_7$ at 77K from calcium doping of grain boundaries. *Nature* **407**, 162–164 (2000).
55. Daniels, G. A., Gurevich, A. & Larbalestier, D. C. Improved strong magnetic field performance of low angle grain boundaries of calcium and oxygen overdoped $\text{YBa}_2\text{Cu}_3\text{O}_7$. *Appl. Phys. Lett.* **77**, 3251–3253 (2000).
56. Feldmann, D. M. et al. Influence of nickel substrate grain structure on $\text{YBa}_2\text{Cu}_3\text{O}_7$ supercurrent connectivity in deformation-textured coated conductors. *Appl. Phys. Lett.* **77**, 2906–2908 (2000).
57. Feldmann, D. M. et al. Inter and intra-grain transport measurements in $\text{YBa}_2\text{Cu}_3\text{O}_{7-x}$ deformation textured coated conductors. *Appl. Phys. Lett.* (in the press).
58. Gurevich, A. & Cooley, L. D. Anisotropic flux pinning in a network of planar defects. *Phys. Rev. B* **50**, 13563–13576 (1994).
59. Amrein, T., Schultz, L., Kabius, B. & Urban, K. Orientation dependence of grain boundary critical current in high- T_c bicrystals. *Phys. Rev. B* **51**, 6792 (1995).
60. Attenberger, A. *Elektrische transportmessungen über definierte Korngrenzen-Strukturen in epitaktischen Bi-2223 Schichten*. Thesis, Technische Universität Dresden (2000).
61. Tsay, Y. N. et al. Transport properties of $\text{Bi}_2\text{Sr}_2\text{CaCu}_2\text{O}_{8-x}$ bicrystals with [001] tilt grain boundaries. *IEEE Trans. Appl. Supercond.* **9**, 1622–1625 (1999).
62. Polyanskii, A. A. et al. Magneto-optical study of flux penetration and critical current densities in [001] tilt $\text{YBa}_2\text{Cu}_3\text{O}_7$ thin-film bicrystals. *Phys. Rev. B* **53**, 8687–8697 (1996).
63. Polyanskii, A. A. et al. Magneto-optical studies of the uniform critical state in bulk MgB_2 . *Supercond. Sci. Technol.* **14**, 811–815 (2001).
64. Larbalestier, D. C. et al. Strongly linked current flow in polycrystalline forms of the superconducting MgB_2 . *Nature* **410**, 186–189 (2001).
65. Chu, S. Y. & McHenry, M. E. Growth and characterization of $(\text{Bi,Pb})_2\text{Sr}_2\text{Ca}_2\text{Cu}_3\text{O}_x$ single crystals. *J. Mat. Res.* **13**, 589–595 (1998).
66. Malozemoff, A. P. et al. HTS wires at commercial performance levels. *IEEE Trans. Appl. Supercond.* **9**, 2469–2473 (1999).
67. Huang, Y. B. et al. Progress in Bi-2223 tape performance. *Adv. Cryo. Eng.* (in the press).
68. Pashitski, A. E., Polyanskii, A., Gurevich, A., Parrell, J. A. & Larbalestier, D. C. Magnetic granularity, percolation and preferential current flow in a silver-sheathed $\text{Bi}_{1.8}\text{Pb}_{0.2}\text{Sr}_2\text{Ca}_2\text{Cu}_3\text{O}_{8+x}$ tape. *Physica C* **246**, 133 (1995).
69. Holzapfel, B. et al. Grain boundary networks in Y123 coated conductors. Formation, properties and simulation. *IEEE Trans. Appl. Supercond.* **11**, 3872–3875 (2001).
70. Rutter, N. & Glowacki, B. Modelling of orientation relations in 2-D percolative systems of buffered metallic substrates for coated conductors. *IEEE Trans. Appl. Supercond.* **11**, 2730–2733 (2001).
71. Rupich, M. W. et al. Low cost Y-Ba-Cu-O coated conductors. *IEEE Trans. Appl. Supercond.* **11**, 2927–2930 (2001).
72. Maeda, T. et al. $\text{YBa}_2\text{Cu}_3\text{O}_x$ thick films grown on textured metal substrates by liquid-phase epitaxy process. *IEEE Trans. Appl. Supercond.* **11**, 2931–2934 (2001).
73. Solovoyov, V. et al. Ex situ post-deposition processing for large area $\text{YBa}_2\text{Cu}_3\text{O}_x$ films and coated tapes. *IEEE Trans. Appl. Supercond.* **11**, 2939–2943 (2001).
74. Holesinger, T. et al. A comparison of buffer layer architectures on continuously processed YBCO coated conductors based on the IBAD YSZ process. *IEEE Trans. Appl. Supercond.* **11**, 3359–3364 (2001).
75. Gurevich, A. Thermal instability near planar defects in superconductors. *Appl. Phys. Lett.* **78**, 1891–1893 (2001).
76. Gurevich, A. & Friesen, M. Nonlinear current flow in superconductors with planar obstacles. *Phys. Rev. B* **62**, 4004–2025 (2000).
77. Friesen, M. & Gurevich, A. Nonlinear current flow in superconductors with restricted geometries. *Phys. Rev. B* **63**, 064521–1–26 (2001).
78. Glowacki, B. A. et al. Superconductivity of powder-in-tube MgB_2 wires. *Supercond. Sci. Technol.* **14**, 193–199 (2001).
79. Suo, H. L. et al. High transport and inductive critical currents in dense Fe- and Ni-clad MgB_2 tapes using fine powder. *Adv. Cryo. Eng.* (in the press).
80. Grasso, G. et al. Preparation and properties of unsintered MgB_2 superconducting tapes. *Adv. Cryo. Eng.* (in the press).
81. Kumakura, H., Matsumoto, A., Fujii, H., Takano, Y. & Togano, K. Fabrication and superconducting properties of powder-in-tube processed MgB_2 tapes and wires. *Adv. Cryo. Eng.* (in the press).
82. Grasso, G. et al. Large transport critical currents in unsintered MgB_2 superconducting tapes. *Appl. Phys. Lett.* **79**, 230–232 (2001).
83. Malozemoff, A. P. et al. D. F. Low-cost YBCO coated conductor technology. *Supercond. Sci. Technol.* **13**, 473–476 (2000).
84. Pyon, T. & Gregory, E. Nb_3Sn conductors for high energy physics and fusion applications. *IEEE Trans. Appl. Supercond.* **11**, 3688–3691 (2001).
85. Field, M., Hentges, R., Parrell, J., Zhang, Y. & Hong, S. Progress with Nb_3Sn conductors at Oxford Instruments—Superconducting Technology. *IEEE Trans. Appl. Supercond.* **11**, 3692–3695 (2001).
86. Wang, W. G. et al. Engineering critical current density improvement in Ag-Bi-2223 tapes. *IEEE Trans. Appl. Supercond.* **11**, 2983–2986 (2001).
87. Masur, L. et al. Long length manufacturing of high performance BSCCO-2223 tape for the Detroit-Edison Power Cable Project. *IEEE Trans. Appl. Supercond.* **11**, 3256–3260 (2001).
88. Hayashi, K. et al. Development of Ag-sheathed Bi2223 superconducting wires and their applications. *IEEE Trans. Appl. Supercond.* **11**, 3281–3284 (2001).
89. Jiang, J. et al. Through-process study of factors controlling the critical current density of Ag-sheathed $(\text{Bi,Pb})_2\text{Sr}_2\text{Ca}_2\text{Cu}_3\text{O}_x$ tapes. *Supercond. Sci. Technol.* **14**, 548–556 (2001).
90. Rikel, M. O. et al. Overpressure processing Bi2223/Ag tapes. *IEEE Trans. Appl. Supercond.* **11**, 3026–3029 (2001).
91. Flukiger, R. et al. Phase formation in Bi,Pb(2223) tapes. *IEEE Trans. Appl. Supercond.* **11**, 3393–3398 (2001).
92. Wang, C. P., Do, K. B., Beasley, M. R., Geballe, T. H. & Hammond, R. H. Deposition of in-plane textured MgO on amorphous Si_3N_4 substrates by ion-beam-assisted deposition and comparison with ion-beam-assisted deposition yttria-stabilized-zirconia. *Appl. Phys. Lett.* **71**, 2955–2957 (1997).
93. Groves, J. R. et al. High critical current density $\text{YBa}_2\text{Cu}_3\text{O}_{7-x}$ thick films using ion beam assisted deposition MgO bi-axially oriented template layers on nickel-based superalloy substrates. *J. Mater. Res.* **16**, 2175–2178 (2001).
94. Hasegawa, K. et al. in *Proc. 9th Int. Symp. Supercond. (IIS'96)* Vol. 2, 745–748 (Springer, Tokyo 1997).
95. Metzger, R. Superconducting tapes using ISD buffer layers produced by evaporation of MgO or reactive evaporation of magnesium. *IEEE Trans. Appl. Supercond.* **11**, 2826–2829 (2001).
96. Reade, R. P., Berdhal, P., Russo, R. E. & Garrison, S. M. Laser deposition of biaxially textured yttria-stabilized zirconia buffer layers on polycrystalline metallic alloys for high critical current Y-Ba-Cu-O thin films. *Appl. Phys. Lett.* **61**, 2231–2233 (1992).
97. Ivanov, Z. G. et al. Weak links and dc SQUIDS on artificial nonsymmetric grain boundaries in $\text{YBa}_2\text{Cu}_3\text{O}_{7-x}$. *Appl. Phys. Lett.* **59**, 3030 (1991).
98. Char, K. et al. Bi-epitaxial grain-boundary junctions in $\text{YBa}_2\text{Cu}_3\text{O}_7$. *Appl. Phys. Lett.* **59**, 733–735 (1991).

Acknowledgements

The authors are grateful to many colleagues for discussions and collaborations. In particular, we thank our present and former Madison colleagues S. Babcock, X. Cai, L. Cooley, G. Daniels, C.-B. Eom, E. Hellstrom, J. Jiang, P. Lee, J. Reeves, M. Rikel and X. Song, and colleagues within the Wire Development Group, especially B. Riley (AMSC), V. Maroni (ANL) and T. Holesinger (LANL). Recent collaborations on coated conductors with T. Peterson and P. Barnes (AFRL), R. Feenstra (ORNL) and D. Verebelyi (AMSC) have been particularly helpful. R. Blaugher (NREL), G. Grasso (Genoa) and J. Mannhart (Augsburg) supplied material for the tables and figures. Finally we thank the Air Force Office of Scientific Research, the Department of Energy and the National Science Foundation for support.

Supported by Materials Research Society (MRS) and
European Materials Research Society (E-MRS)

a First Materials Science Forum on



MATERIALS SCIENCE FOR FUTURE SUSTAINABLE TECH- NOLOGIES

will take place at the

University of Augsburg, Germany

on September: 17-20, 2002

Organizing Institutions: AMU (Application Center for Materials and Environmental Research,
Institute of Physics, University of Augsburg)

WZU (Environmental Science Center, University of Augsburg)

WEC (World Environment Center)

IMU (Institute of Management & Environment)

BI (Bayern Innovativ) (Logistics)

Topic: An informative and at the same time strategic discussion meeting between **scientists, high-ranking industrial leaders, and politicians** is planned to **highlight the crucial role of materials science for the development and the implementation of sustainable technologies**. Based on the present state of affairs and the problems of a partly devastating utilization of resources, products, and fuel, new approaches and strategies for the design and production of sustainable and appropriate functional materials and efficient processes will be discussed. The aim of the meeting is the exchange of ideas and experiences on environmentally benign synthesis routes and utilization of materials by efficient and economically feasible process technologies.

The discussions during this meeting will be inspired by the newest political decisions of the Rio+10 world conference held in Johannesburg two weeks before.

Among others, the conference will comprise the following topics:

Fundamentals: Evaluation of the current state of affairs

Needs: Both industrial and human needs

Strategies: The fuel-energy enigma, functional composite materials, ceramics, catalytic materials, semiconductors, metals, glasses, carbon modifications, biomaterials and biomimetic materials, waste, and life-cycle assessment

Conclusions will be developed during a **panel discussion** with the goal that the participants of the conference reach a consensus on recommendations **for adequate scientific, economic, and political programs in materials science leading to sustainable technologies** to be launched by the individual governments. An **information management system** will be installed on the newest development in the area of materials science for sustainable technologies.

Invited speakers include: Hartmut Grassl, Max Planck Hamburg, Germany, advisor to the chancellor

Louis Schlappbach, EMPA, Switzerland

Mrityunjay Singh, NASA, USA

For Information:

Materials Research Society
506 Keystone Drive
Warrendale PA 15086-7573 USA
Tel: 724-779-3003
Fax: 724-779-8313
E-mail: info@mrs.org
http://www.mrs.org

European Materials
Research Society
BP20
67037 Strasbourg Cedex 2
France
Tel: 33 (0)3 88 10 63 73
Fax: 33 (0)3 88 10 63 43
E-mail: emrs@phase.c-strasbourg.fr
http://www-emrs.c-strasbourg.fr

International Advisory Committee members include:

Hartmut Grassl, Max Planck Hamburg, Germany,
advisor to the chancellor

C.N.R. Rao, IIS, India
Armin Reller, Uni Augsburg, Germany
Mrityunjay Singh, NASA, USA
John Meurig Thomas, Uni Cambridge, UK
Bernard Tramier, Total Fina Elf, France
Henrik Topsoe, Haldor-Topsoe, Denmark
James G. Veras, WEC, USA

Technical Organizing Committee:

Armin Reller, WZU (Chair)
Wolfgang Biegel, AMU
Siegfried Horn, Uni Augsburg
Josef Nassauer, Bayern Innovativ
Bernd Stritzker, AMU
Guido Weissmann, AMU

International Scientific Committee members include:

Bernd Stritzker, Uni Augsburg (Chair)
Johannes G. Bednorz, IBM Rüschlikon, Switzerland
Delcie Durham, National Science Foundation, USA
Jose M. Gonzalez-Calbet, Uni Madrid, Spain
Siegfried Horn, Physics Uni Augsburg, Germany
Marc Ledoux, Uni Strasbourg, France
Toni Marechaux, National Materials Advisory Board, USA
Josef Nassauer, Bayern Innovativ, Nürnberg, Germany
Bernd Wagner, Uni Augsburg
Jay Lee, University of Wisconsin, USA
Robert Pfahl, Motorola, USA

Contacts

Publisher: Fabien Savenay
Editor: Paul Smaglik
Sales Director: Ben Crowe

European Head Office, London

The Macmillan Building
4 Crinan Street
London N1 9XW, UK
Tel +44 (0) 20 7843 4961
Fax +44 (0) 20 7843 4996
e-mail: naturejobs@nature.com

Group European Manager: Leonie Welss (4954)

European Manager:
Nevin Bayoumi (4978)

UK/ RoW/ Ireland:

Matt Powell (4953), Ben Corp (4974),
Andy Douglas (4975)

Holland/ Italy: Nevin Bayoumi (4978)

Scandinavia: Sille Opstrup (4994)

Spain/ Portugal: Leonie Welss
(4954)

Production Manager: Billie Franklin

To send films and materials use
London address above.
Tel +44 (0) 20 7843 4814
Fax +44 (0) 20 7843 4996
e-mail: naturejobs@nature.com

International

Advertising Coordinator:

Laura Pearson (4977)

Naturejobs web development:

Tom Hancock

Naturejobs online production:

Ben Lund

European Satellite Offices

France/ Belgium:

Christine Niox-Chateau
Tel + 33 (0) 1 43 20 16 51
Fax + 33 (0) 1 43 20 51 52
e-mail: c.nioxchateau@nature.com

Germany/ Austria/ Switzerland:

Patrick Phelan/ Kate Turner
Tel + 49 89 54 90 57 11/-2
Fax + 49 89 54 90 57 20
e-mails: p.phelan@nature.com
k.turner@nature.com

US Head Office, New York

345 Park Avenue South,
10th Floor, New York, NY 10010-1707
Tel +1 800 989 7718
Fax +1 800 989 7103
e-mail: naturejobs@natureny.com

US Sales Director: Ben Crowe

US Sales Manager: Peyton Mason

US Advertising Coordinator:

Corissa Salzman

Japan Head Office, Tokyo

MG Ichigaya Building (5F),
19-1 Haraikatamachi,
Shinjuku-ku,
Tokyo 162-0841
Tel +81 3 3267 8751
Fax +81 3 3267 8746
e-mail: k.cowan@naturejpn.com
Japan Manager: Kate Cowan

naturejobs

Radiating good will

It's been a good few months for synchrotrons. The Swiss Light Source recently began testing its beamline at the Paul Scherrer Institute in Villigen. A United Nations body this month formally approved the plans for SESAME, a synchrotron facility to be based in Jordan. And the US National Institutes of Health announced plans to add three beamlines to Argonne National Laboratory's Advanced Photon Source.

All these developments, in addition to existing plans to build two more synchrotrons — one in Britain, the other in France — are welcome news to two groups of scientists. More beamlines will make life easier for structural biologists, who use the high-energy radiation to decipher the three-dimensional structures of molecules. More synchrotrons also offer the prospect of more jobs for the physicists and engineers needed to operate them.

For structural biologists, an increasing demand for access to beamlines — mainly because of the rising profile of proteomics — has meant that most researchers have to wait their turn in a growing queue, and often travel a long way, to do their experiments.

But additional facilities, by themselves, only address half the problem. Without experienced engineers and physicists to run them, the machines will not be able to serve their clients as well as they might. But specialized staff are not easy to find — skills have often been developed only by working at synchrotrons for years.

The new facilities will need to find ways of recruiting existing specialists and, hopefully, should plan to train new ones and find ways of retaining both. Perhaps a start would be to ensure that the technicians who run the machines receive due credit when new structures are solved.

Paul Smaglik
Naturejobs editor



Contents

MOVERS

Academic takes the plunge in industry; Sales manager turns his back on stability; Optical society gets a new vice-president; and more [Back page](#)

WWW.NATUREJOBS.COM

Career centre
Information on the scientific
job market

FOCUS

SPOTLIGHT

RECRUITMENT

SCIENTIFIC ANNOUNCEMENTS

SCIENTIFIC EVENTS

PHARMACEUTICALS

During his 25 years in academia, **Robert Lenox** turned down several jobs in industry. But now he has taken the plunge and become head of the central-nervous-system disease group at Aventis. Lenox left his endowed professorship at the University of Pennsylvania in part because he recognized a shift in the drug-discovery process. In recent years, he says, pharmaceutical research concentrated more on combinatorial chemistry. But now rational drug design is in vogue because "you can find mechanisms that are critical to the disease process", he says.

Throughout his academic career, Lenox was fascinated by the molecular mechanisms that underlie complex behavioural conditions such as bipolar disorder. He always thought that, at some point, he would like to be in a position to help search for better treatments for these conditions.

Lenox had anticipated some cultural differences as a result of his move — more resources, but less individual autonomy — but he now realizes that autonomy can also mean little or no collaboration. Since joining Aventis, he has been impressed at the extent to which his new colleagues cooperate on multiple projects. "What particularly struck me is that the scientists with whom I deal really do want to make an impact," Lenox says. "They want to discover drugs that can make a difference."

BIOTECHNOLOGY

Andrew Webb, who left Amersham Biosciences in England last month to become sales and marketing director for DxS, a biotech start-up based in Manchester, UK, has travelled a merger-filled path from bench to sales, and from pharmaceuticals to biotech. "You get quite relaxed about it after a while," he says. "Each time I've fallen on my feet." After university, Webb worked in drug metabolism with SmithKline & French, but took voluntary redundancy when the company closed its preclinical operations in 1989. He joined Beecham Pharmaceuticals and, ironically, six months after he left SmithKline, the company merged with Beecham and he found himself working for the same people again.

Lenox later joined Amersham International's life sciences division in a technical support position, and after the business merged with Pharmacia Biotech, became UK industry sales director and eventually European sales director. He is pleased that his new position with DxS will encompass many of the roles he has had in what he

describes as six jobs in 11 years, but jokes that he has left his quest for stability behind. "Amersham Biosciences is about as secure a position I could've had," he says.

After 25 years with Bristol-Myers Squibb, **Michael Howerton** this autumn joined ImClone Systems, a New York-based biotech company, as vice-president, business development. Howerton will oversee licensing arrangements and product acquisitions. At his previous company, he served as vice-president, financial analysis and assistant controller, having spent the previous eight years as vice-president of corporate development, a role similar to his present one at ImClone.

COMPUTER SCIENCE

This month, **Danny Powell** became executive director of the National Center for Supercomputing Applications (NCSA) at the University of Illinois at Urbana-Champaign. Powell most recently served as associate director of the Los Alamos Computer Science Institute and as the associate director for the Center for High Performance Software Research at Rice University. Powell will manage the day-to-day operations of the NCSA and work to link it with the National Computational Science Alliance. Powell succeeds **Jim Bottom**, who left the NCSA in the summer to become vice-president for information technology and chief information officer at Purdue University in Indiana.

PHYSICS

Peter Knight, who recently became head of the physics department at Imperial College in London, was appointed vice-president of the Optical Society of America (OSA) last month. This is the first time someone living outside North America has held the position. Although the OSA's headquarters are in Washington, Knight points out that a third of the society's 14,000 members work outside the United States, and that its premier journal, *Optics Letters*, currently publishes more papers by Europeans than Americans. Knight will serve as vice-president next year, president-elect in 2003 and president in 2004.

ERRATUM

In Movers in the 18 October issue of *Naturejobs*, John Kelly is incorrectly described as overseeing a \$350 million annual research programme. He will in fact oversee a \$350 million development of a new research site for the University of California, San Francisco, at Mission Bay.



Robert Lenox



Andrew Webb



Peter Knight

CONTACTING US AT MOVERS

Please send any information on new appointments or job moves to: naturejobseditor@naturedc.com.